Analysing social media data using sentiment analysis in relation to public order

James Baldwin

A thesis submitted in partial fulfilment of the requirements of
Sheffield Hallam University
for the Degree of Doctor of Philosophy

September 2021

Candidate Declaration

I hereby declare that:

1. I have not been enrolled for another award of the University, or other academic or professional organisation, whilst undertaking my research degree.

2. None of the material contained in the thesis has been used in any other submission for an academic award.

3. I am aware of and understand the University's policy on plagiarism and certify that this thesis is my own work. The use of all published or other sources of material consulted have been properly and fully acknowledged.

4. The work undertaken towards the thesis has been conducted in accordance with the SHU Principles of Integrity in Research and the SHU Research Ethics Policy.

5. The word count of the thesis is 85,000.

Name	*Analysing social media data using sentiment analysis in relation to public order*
Date	*September 2021*
Award	*PhD*
Faculty	*Business, Technology and Engineering*
Director(s) of Studies	*Jotham Gaudoin*

Table of Contents

Table of Figures

List of Tables

Acknowledgements

I would like to give special thanks and express my gratitude to my PhD supervisory team, Jotham Gaudoin, Teresa Brunsdon, Laurie Hirsch and Keith Burley, for there continuous encouragement, guidance, advice, ideas and reassurance throughout my whole PhD process.

I would like to express my appreciation to Sheffield Hallam University and Industry and Innovation Research Institute (I2RI) for support and the scholarship opportunity I was given to learn and develop my skills in academia.

I am really thankful for my friends constant encouragement throughout my academic studies. I am very grateful for my family's patience, endless love and support throughout my PhD degree and for all the amazing things in helping me towards my achievements. My family have given me the opportunity, motivation and determination to pursue my ambitions. A special thank you to my wife Sijia who encouraged and supported me in fulfilling my PhD journey. Thank you all for always believing in me when I doubted myself at times it has made all the difference.

Abstract

The research aim is to analyse social media data using sentiment analysis in relation to public order. A sentiment can be expressed in a thought, opinion or attitude that is mainly based on emotion instead of reason. (SA) Sentiment Analysis studies the opinions, sentiments and emotions expressed at sentence or document level. SA extracts text which is identified and classified as opinions or emotions that aim to support a decision-making process through the analysis of text. SA identifies and measures whether the text being analysed is positive, negative or neutral in relation to an entity, such as people, organisation, event, location, or a topic. As the adoption of ubiquitous technology increases and the population on social media continues to grow with the speed of responsiveness of the users expressing their political, economic or religious views on Twitter or Facebook, the posts become valuable sources of public opinion. This can be seen as an important commodity to be used to infer public opinions for social studies or marketing.

The research suggests the police have found it difficult to adapt their existing model to the changing nature of public events and handling of acceleration towards technology and social media. The scalability and volume of data has made it increasingly hard for the police to manage, monitor and make use of intelligence emerging from social media to maintain the peace. To address this gap, the investigation will evaluate whether SA can enhance the analysis of social media in the context of public (dis)order events. This may help to improve the police's decision-making process and reduce complexity to increase public safety. There are specific and generalised ways that SA can support the police, but this research might focus on a specific case. To meet the aim, the research proposes to use a SA model, data mining tools and techniques to analyse the relevant data extracted from social media. The project will use an adapted social media lifecycle as a methodological approach. Past events involving public order and the police will be evaluated to develop relevant methodology and provide appropriate recommendations to the technical community on ways to use SA for future applications of social media.

In the project it adopted a hybrid approach which consists of a dictionary, machine learning and gold standard approaches. As result, the machine learning of dictionaries and manual classification results proved to show the strongest output based on precision, recall and F1 measure when compared to the machine learning of tweets and manual classification. The change point analysis helped to identify significant points in the timeline of tweets for the event which correlated to the physical event. However, there were some inaccuracies on the allocated points of change, as deemed insignificant based on news media and low volume of tweets. Future work is required to understand the reasons behind the allocation change points and possible use of alternative methods to help extract further insights that could not be explored in this project.

The study makes a series of contributions to knowledge. First, to the creation of a keywords for public order events due to none being publicly available. Second, is to build towards a model to predict what may happen in public order events with the application of dictionary, machine learning and creation of gold standard in the realm of sentiment analysis. Third, the technical contribution to sentiment analysis community to help provide future recommendations to potentially enhance their framework and what areas require further research in the area. Fourth, is the development of social media lifecycle methodology, which has been tested in this project.

1 Introduction

1.1 Aim

To analyse social media data with the use of sentiment analysis in relation to public order events.

Research questions:

1) Can we determine changes in sentiment over time as recorded in social media data emerging from public order events?
2) Is there any current sentiment dictionary that is effective at determining the level of sentiment accurately for public order events?
3) Is a dictionaries and manual classification machine learning approach more effective than the input of tweets and manual classification approach?

1.2 Objectives

1) Investigate the historical and ongoing development of police practice in the use of social media and complementary technologies used in the management of public disorder.
2) Evaluate the effectiveness of various approaches to sentiment analysis using social media data from public order events.
3) Identify and determine both the suitability and relevance of social media platforms
4) Decide on which suite of research methods and instruments to implement for data extraction.
5) Identify and collect thematic range of data and use data mining, text mining and sentiment analysis and other tools and techniques to analyse it.
6) Analyse the data collected with identified tools and techniques
7) Develop relevant methodology and provide appropriate evaluation and recommendations to the technical community on how to use SA for future applications of social media.

1.3 Elaboration

Since 2011, interest has grown in social media from both the academic and industrial perspectives (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018). For example, Law Enforcement Agencies substantially increased their usage of social media data, with policy changes being implemented to adapt to social media and its possible uses after the 2011 London riots occurred (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018). This interest has to some extent been driven by the rapid increase in usage of social media networks and of internet accessibility; the internet was used daily or almost daily by 82% (41.8 million) of UK adults, compared with 78% (39.3 million) in 2015 and 35% (16.2 million) in 2006 (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018). Organisations now have social media teams to monitor events and actively release information, quickly reacting to situations of widespread interest (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018).

The increase of technology and population on social continues to grow at rapid speed of responsiveness where users express their opinion on different topics, such as economic and political views on social media platforms, such as Twitter, Facebook and YouTube which are important source of public opinion. The research conducted indicates that the police seem to have difficulty with their model to adapt to the evolution of public events and handling of increase in speed towards technology and social media. The police practice in the United Kingdom (UK) when approaching social media analysis in a public order event is seemingly more a manual process where members of the police use different mobile devices to try to analyse what's going on within these social media platforms. This approach by the law enforcement agency poses an operational problem that is resource intensive when trying to disseminate information, build intelligence and prevent indignation spreading when trying to adapt situation on-demand (HMIC, 2011b).

The research aims to enhance the analysis of social media data using sentiment analysis in relation to public order. This requires an investigation into social media, how it is stored (big data issues), collected (data extraction) analysed (text mining, sentiment analysis) and then disseminated to the police. The research project could have made use of focus groups, but to capture the wider opinion on public order events with the use of sentiment analysis to detect the affection of tweets to predict what may happen next in an event. The research focuses on Twitter as it's the most open platform and is widely used by the research community (Sloan & Quan-Haase, 2016). The advantage of Twitter over its competitors are due to the short character limit per tweet that encourages Twitter users to provide live updates, at any time, in any location and on any device (Sloan & Quan-Haase, 2016). Furthermore, the retweet capability helps to further disseminate a message from various movements that occur, increasing awareness of a given demonstration.

To help achieve goal of the project, public disorder events will need to be identified and suitable data extracted from the Twitter circulating these events, to discover what might be deemed the most preferable analytic processes to apply. The project will help to build towards a model that can predict what may happen based using a keywords list to find the critical elements of information within social media. This model will help automate the process that can be used a wider context, but in this case, it will look to provide greater support for policing of public order. For instance, this model may also support the police to gain greater insight into their community to prevent tension and conflict, bringing greater cohesion within the local community.

1.4 Structure of The Thesis

The proceeding sections of the thesis are organised as follows:

- The research domain provides background information on historical, current and future police practice, models of policing public order overtime, police use of social media that includes information sharing, engagement and intelligence based on Twitter, Facebook and YouTube, social media audience, role of social media, organisations and activism, and police usage of text mining, data mining and general to police use of sentiment analysis, which includes a focus on Twitter. Additionally, there is a general use of sentiment analysis covered more widely due to limited research on police usage of sentiment analysis. The literature has helped enhance researchers' knowledge of the different areas listed above, some of which covers elements around methodology. This helps inform choices on the chosen methodology which outlined in the methodology chapter.

- The methodology evaluated and a justified approach selected to be applied to the project, which includes use of social media lifecycle, qualitative, quantitative, and case study approaches. The associated tools and techniques, such as TAGS, R programming and DiscoverText that are used from data acquisition to data analysis. Moreover, social media platforms are identified and compared against each other, where Twitter is justified for use to the project. Furthermore, the pilot study and UK demonstrations datasets will be carefully chosen. The different sentiment analysis approaches are explored and a hybrid method (includes dictionary, machine learning and gold standard) selected for the classification of tweets. The methodological framework set in this section will be followed in the implementation phase.

- The pilot study is conducted on the Baltimore dataset, which the results of the process and analysis will help to inform whether any changes are required to the methodological framework when analysing UK demonstrations.

- The exploration data analysis chapter explores the UK demonstration datasets from meta data, language, retweets to bad data, discusses preparation of data to coding and cleansing of data and devises relevant analysis techniques from both different dictionaries, algorithms, change point detection and evaluation methods e.g., precision, recall and F1 measure to gain useful insights. The initial findings for each UK demonstration to provide some insights into top hashtags, peak and troughs for each event over time with a general overview by day and specific views by day and hour, identification of keywords with use of word clouds, popularity of mobile devices and lexical diversity. The insight gained from this chapter will help form a greater understanding of the results from the sentiment analysis phase.

- The sentiment analysis and change point analysis results chapter discusses the data extraction, coding of datasets, pre-processing of data, initial exploration of the lexicon results over time, evaluation of sentiment analysis which a sample of data undergone manual classification, inter agreement between different manual classifiers is explored. Furthermore, the results for dictionary, machine learning and gold standard approaches are analysed based on their precision, recall and F1 measure. Moreover, the change point results element explores the sentiment classification before, during and after the event and is aligned with news media reports and tweets to support the reasons for the change through the event. Change point techniques, such as BinSeg are used to identify significant points of change based individual sentiment categories, such as negative, neutral and positive. The results of specific algorithms, such as Max Entropy and Naïve Bayes are displayed over a period of time based on their predictions. These results from the sentiment analysis stage will provide indication on the strength of dictionaries and algorithms results, and whether any changes are required to any of the framework to further enhance the output.

- The final chapter evaluates the key stages of the project, such as pilot study, data exploration, hybrid approach and the change point analysis results to identify strengths and weaknesses with a series of improvements put forward. The next element is to evaluate whether the aim, research questions, objectives and deliverable have been achieved, then a a series of recommendations based on the evaluation of the project. The final part is a conclusion of the project's main points will be outlined.

Lastly, the appendices are attached at the end of the document as supplementary information. This includes the ethical approval and additional results from the analysis that are not included in the main findings of the project. In chapter 2 the research domain will be explored in great depth to understand the police, public

order, social media and the technicalities of sentiment analysis and machine learning that will support the project.

2 The Research Domain

This research will draw together relevant background information relating to the objectives, and these are now each discussed below to understand each of the themes and any gaps identified in the literature.

2.1 Historical, Current and Future Practices within the Police Force

The Law Enforcement Agency's (LEA) operations have transformed substantially over the last 20 years, and will continually change, as depicted below in Figure 2.1. This timeline indicates the change from 2015 in April the Association of Chief Police Officers (ACPO) to National Police Chiefs' Council (NPCC). The police are adept at reacting to the here and now, busy focusing on demand. Research seems to indicate the need of taking greater advantage from the considerable corpus of data available patterns must be detected which can possibly be used to influence day-to-day operational strategies and at longer term crime initiatives (Bullock, 2015; House of Commons, 2018; NPCC, 2015a). The reasons for the police's structural reform is led by the formation and rapid development of the Internet and its social applications, increase in personal mobility and migration and fragmentation of society. This has seen change in the type of criminal opportunities, threats and risks, leading to increased public demand on security and order (Castells, 2009; Chui, 2012; IPC, 2013). One other key factor in these reforms results from Governmental budget cuts to the police (Pickles, 2015; IPC, 2013). To make the police force more efficient and effective, the NPCC will be *"focussing on operational delivery and developing national approaches on issues such as finance, technology and human resources"* (Bullock, 2015; NPCC, 2015a) by working closely with the College of Policing that is responsible for developing professional standards. It is apparent that the police will constantly need to adapt their practice to the changing environment (IPC, 2013), but will need to be based upon Sir Robert Peel's nine policing principles (Durham Constabulary, 2017) when serving and protecting the public (Home Office, 2012; IPC, 2015). This research suggests the police need to revise their capability to adapt to socio-economic transformations in society.

Current police practice is moving towards a nationwide approach to help reduce crime and provide public safety, but with a view of a joint operational response to the most serious and strategic threats (ACPO, 2015). The NPCC (2015) Chair, Sarah Norton, argues that there are two challenges in changing requirements and cost pressure on how we re-imagine policing in the UK. Her Majesty's Inspectorate of Constabulary (HMIC) (2015) has echoed for a change in police requirements due to cost and changes in public demand.

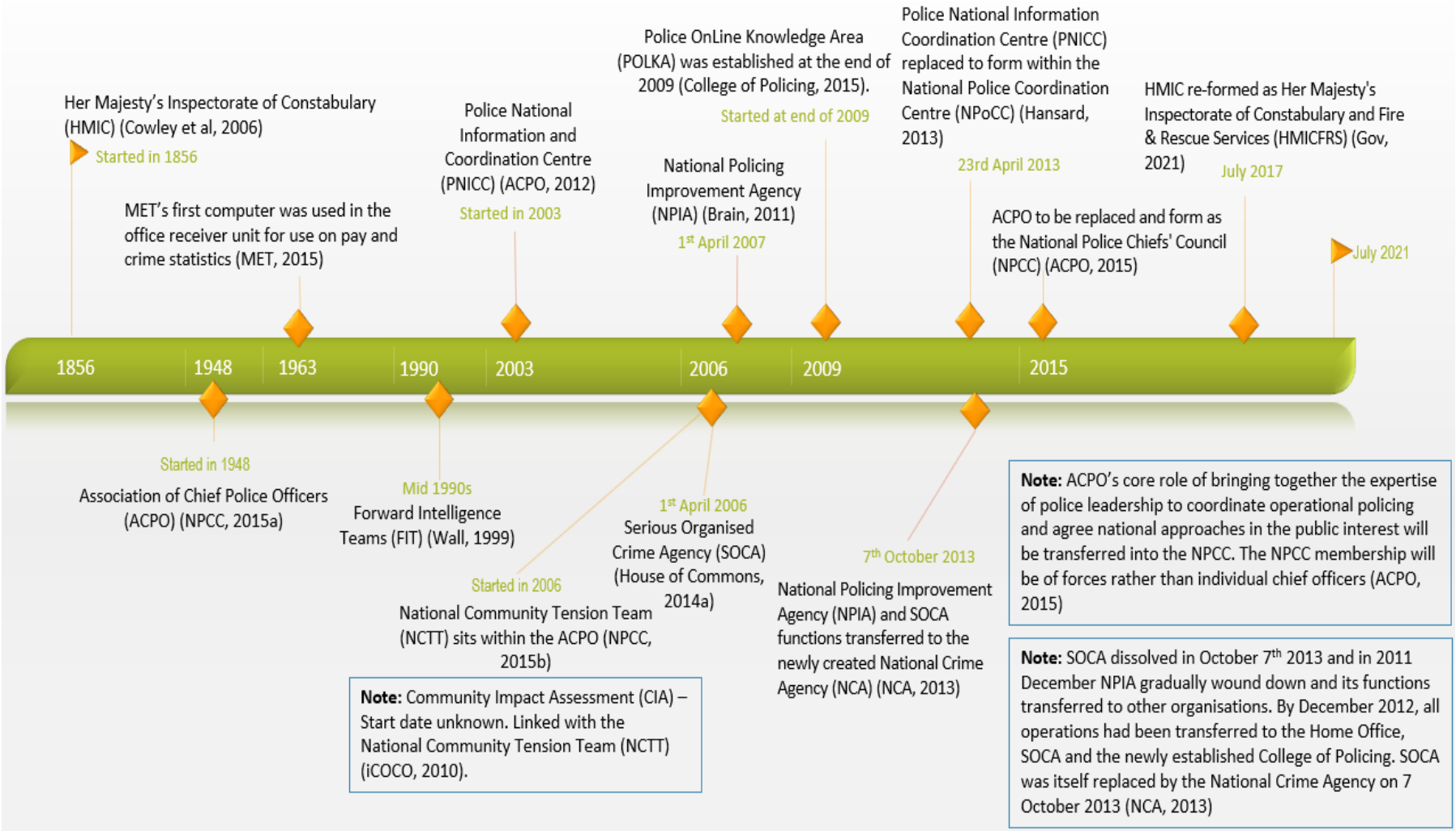

Figure 2.1 Limited view of police historical timeline (Baldwin, 2015)

HMIC (2015) and House of Commons (2018) shows there is a major shift towards reports of increasing mental health calls and crimes being committed on-line. Similarly, there have been increased incidence of electronic fraud and cybercrime on social media and other web platforms and raised need for counter-terrorism measures (Bullock, 2015; HMIC, 2015). Police forces are constantly being challenged through budget cuts, evolving security landscape (notably ever-increasing cybercrime) and new emergence of digital technologies (HMIC, 2015; Metropolitan Police Service, 2014; NPCC, 2015). Police organisations are seemingly overcoming these issues by consideration of improved utilisation of technologies and operational, organisational and cultural changes to deliver an effective, legitimate and committed service to the public (HMIC, 2015; Metropolitan Police Service, 2014 & 2021). The Metropolitan Police Service (MPS). For example, are investing approximately £200 million to upgrade Information and Communications Technology (ICT) infrastructure (including advanced data analysis tools to predict crime trends and hot-spots to increase crime prevention) and introduce the use of wearable technology (such as a public order helmet cameras) (Metropolitan Police Service, 2014 & 2021). This technology may help to predict crimes before these are committed. Kent Police, for instance, are adopting a pre-emptive approach, in which all the crime information is inputted into a predicative system (which combines historic data with psychological assessment features to identify level of risk) (Metropolitan Police Service, 2014 & 2021). Subsequently, a database of information is establish to permit identification of geographical hotspots and ensure correct allocation of policing, so the police can identify hotspots on the map and make sure officers are in the right place (Watts, 2013).

Despite such examples of progress, Dearden (2017) notes that the report produced by Royal United Services Institute for Defence and Security Studies (RUSI) in 2017 highlights many police forces lack the capability to use the wealth of data as a result of cultural barriers, lack of legal frameworks and ethical concerns. Predictive technology should not be viewed as a means by which to replace officers' practical skills 'on the street' but considered as enhancement through the provision of additional information. An appreciation of the role of technology must be emphasised as staff have shown genuine concerns about possible minimisation of job roles; it is essential reassurances are provided to prevent negative attitudes towards integration of technology (Dearden, 2017; Watts, 2013). There are many factors that could affect the change as there is potential unreliability of emerging technology, the limited progress of consolidating the fragmented databases and disparity of the police forces' legacy systems (Dearden, 2017; Watts, 2013). Expenditure on technology is justified if its capability is utilised in an effective manner (Dearden, 2017; Watts, 2013).

RUSI (2017) document the urgent requirement for an ethical framework and clear national guidance to drive the nationwide procurement of new technology and encourage improved data sharing practices between individual forces and other organisations. Additionally, appropriate training and adequate other resources must

be established to maximise effectiveness of use. This research recommends further progress must be made to effectively use this technology in the future. The Home Office recognises the importance of improving technological capability and is working to address the issues identified by RUSI (2017), *"with more than £1bn invested in national law enforcement digital programmes"* (Dearden, 2017). Big data can be explored to a greater depth to learn from crimes in the past to simulate crimes of the future with the need for an appropriate ethical framework (Babuta, 2017; Kearns & Muir, 2019; Watts, 2013). Information that is already known about the spread of medical diseases on social media may be applied in the creation of a model to predict the spread of public riot to increase public safety. It may permit identification of potential geographical areas in which unrest may occur to permit the use of different strategies to minimise public disorder.

The police understand that through technology it can enhance policing to a good effect (Koper, Lum & Willis, 2014; Gash & Hobbs, 2018). The use of technology is to improve the police's ability to identify and monitor criminals, facilitate the identification of locations and the conditions of a setting and improve speed of response to crimes. Moreover, may help bridge the communication gap between the police and the public, and provide the police with skills to enforce laws in relation to tech advanced technological crimes, such as identity theft and cybercrime (Koper, Lum & Willis, 2014). The implementation of differing technologies can present difficulties and complexities, particularly in relation to technical issues such as end-user functionality problems and user interface issues, cultural resistance and training of the workforce. Simply relying on technology is not the panacea for everything, the police must still continue to work with and build relationships with the wider community, as not all information will be found within Big Data, as recently shown in the political landscape voting in elections (Babuta, 2017). The people and community leaders are vital to gain insights into the community to increase public safety. Technological advancements in surveillance, protective equipment and weapons may assist in reduction of injuries or deaths of officers, suspects and members of the public (Babuta, 2017; Christie, 2021; Metropolitan Police Service, 2014 & 2021). Surveillance capability may prove controversial depending how privacy and security is balanced to ensure public safety (; Babuta, 2017; CDEI, 2020a & 2020b; Christie, 2021; Kearns & Muir, 2019; Metropolitan Police Service, 2014 & 2021).

The use of big data to tackle crime is still in its infancy so it is difficult to evaluate if it has the potential to be truly revolutionary and predict crime, the police and intelligence agencies must embrace the idea that use of data has other functions rather than accelerating task completion and making economic savings (Watts, 2013). In principle, the upgrade in ICT infrastructure and amalgamation of networks is a major step in the advancement of future policing; as with change, there are reasonable doubts surrounding how police forces will take advantage of this new technological capability, alongside the maintenance of normal demand on service of duty despite

further budget cuts (Constable, 2015a; 2015b; 2015c; Janoowalla, 2015). The issues raised about police resources seemingly need to be addressed by relevant staff in the LEA and within government authorities, so it assures optimal service delivery to the public. It is believed that the police face an uncertain future (IPC, 2013; Constable, 2015c) and need to be considered as a greater priority in local and national government, as re-shaping of police may have a profound effect on the public's security within the future.

2.2 Police and Public Order

2.2.1 Policing Public Order

Public order can be characterised as an absence of disorder, as it involves people being considerate, rational, sensible and respecting of others in a public space (OU, 2009). Public and disorder come together when the dynamics of disorder are empowered through the use of public space. For example, if a person on the street is overly exuberant, resulting annoyance may be spread to others. Subsequent police intervention aims to provide order tactically to restore public order (Body-Gendrot, 2014). Public order crimes comprise a range of offences, such as nuisance and consensual offenses and victimless crimes. These may include, for instance, the selling of drugs on a street, being drunken and disorderly, political violence in riot and intimidation against targeted groups or individuals (CPS, 2015; Home Office, 2015; OU, 2009). Victimless crimes are acts that are crimes, such as drugs under law, but have no victim (Walsh & Hemmens, 2011).

Public order policing is referred to the policing of protesters, campaigners and other large gatherings that are either planned or spontaneous (Wakefield & Fleming, 2009). The types of planned events for which there may provide a visible police presence include, for example, demonstrations, trade union picketing, sporting events and concerts. These events tend to involve thousands of people and necessitate the deployment of police officers to maintain order and security (Ho, 2013). It is vital that the policing of these events involves prior knowledge for construction of police tactics and strategies to be appropriate to the gathering (Wakefield & Fleming, 2009).

The UK mainly sees demonstrations that are organised events in-line with the police, rather than a protest occurs on demand due to civil unrest being caused by some level of injustice (HMIC, 2009). Planned events may still involve a degree of unpredictability, as the crowd behaviour may disintegrate into disorder for differing reasons. An anti-austerity march in 2011 in Rome, for example, became a turn for the worse when 30,000 protesters were assaulted by hundreds of Black Bloc (a group who wear black clothing, ski masks or other face concealing items) armed demonstrators (Body-Gendrot, 2014). Spontaneous events, such as gate-crashed parties (with multiple

people overflowing onto public streets), riots (a noisy, violent public disorder caused by a group or crowds of people that are privately acting together in a disruptive and tumultuous manner to purposely disturb the peace), flash mobs (sporadic gathering of large groups of individuals appear in a location for no apparent reason, resulting in senseless behaviour) and illegal raves presents a greater challenge for policing. It is therefore difficult to predict how the situation will evolve making it harder to adapt to that environment (Wakefield & Fleming, 2009). These events, either planned or spontaneous, are challenging for the police, as the balance between the use of police and inaction may escalate to further violence.

Public order is one of the five main threat areas in the Strategic Policing Requirement that sets out to protect the public when disorder occurs (Home Office, 2015). There are six core principles that apply when policing public order operations comprising policing style and tone, communication, use of the National Decision Model, command, proportionate response, capacity and capability (College of Policing, 2013 & 2014). There are three command levels that enforce these principles within the public order command structure that are based on gold, silver and bronze roles (College of Policing, 2013 & 2014; HMIC, 2014; JESIP, 2013), as defined below: -

- **Gold:** A gold commander oversees an incident and is stationed in a control room known as gold command where strategic approaches are developed for the police service to adopt for incident management.
- **Silver:** A silver commander takes strategic direction from a gold commander and creates tactics that are implemented by bronze command.
- **Bronze:** A bronze commander is a member of staff from an emergency service who controls a part of the incident response, implementing silver commander's tactics. In some operations there may be a requirement for a sub-bronze commander role.

These officers in the command are given authority for a role in a specific operation or incident and commanders are to "make decisions, give clear directions and ensure those directions are carried out" (College of Policing, 2014). These commanders must be trained, qualified and operationally competent to perform a role in this command when a public order event arises.

In accordance of the national approach being taken up by the NPCC, there is an agreed national framework for managing the local multi-agency response to emergencies. Command, control and coordination are important concepts in a multi-agency response. Single agencies have often used the gold, silver and bronze control structure. In a large-scale, multi-agency coordination situation, this control structure is convened at strategic, tactical and operational levels.

The police develop communication plans to communicate with the “public, directly or indirectly”, which is crucial to PPO (College of Policing, 2013). These plans include

community engagement that are tailored to each individual event suited to the communities. Neighbourhood and policing implementers draw up plans to liaise with the "local media, key internal/external stakeholders and directly with the public and will support the relationship between the police and the public" in moments of rising tension (College of Policing, 2013). This engagement is planned and developed with the "wider force community engagement plan/ policy" (College of Policing, 2013).

2.2.2 Models of Policing of Public Order

The public order command structure has changed its model of approach to policing of public order to minimise conflict and retain the peace. Several different models to policing of public order are explored below to identify why the police's approach (dis-)order has changed over time:

1) **Escalated Force**
   - The Escalated Force Model employed aggressive tactics to disperse protesters to keep order even when demonstrations were peaceful and within the law; over time the past inherent injustices of the approach have not been widely approved within a democratic context (McPhail and McCarthy, 1998; Waddington, 2011a & 2011b). The LEA's in the UK used this approach in the 1980s for "total control" but shifted towards "Negotiated Management" in the 1990s, which is postulated to show the LEA's have learnt from past mistakes and experiences (Ho, 2013; Waddington, 2011a & 2011b).

2) **Negotiated Management**
   - Negotiated Management Model provided greater respect for the 'right to protest', with a stronger emphasis on negotiation and compromise, reluctance to use force and make arrests, and increased tolerance towards disruption (Stott et al., 2013; Waddington, 2007). This greater level of communication between the police and protesters creates improved control of the event. The police use minimal force to maintain the peace, but the planned approach provides predictability of the event, thus reducing the risk factor (Stott et al., 2013, Waddington, 2007; McPhail and McCarthy, 1998).

3) **Strategic Incapacitation**

   Strategic Incapacitation Model is "defined as excessive controlling of space to isolate and contain potentially disruptive protest, the use of pre-emptive arrest, surveillance and information sharing" (Stott et al., 2013, pg213). Strategic Incapacitation divides public space into different securitised zones, uses surveillance information to incapacitate disorderly protesters, and to manage dissemination and its creation. Strategic Incapacitation was practised by the UK police force in the 2000s, as there was a globalised, non-hierarchal protest movement that was set within the context of technological change and

the growing usage of social media (Waddington, 2011a & 2011b). Strategic Incapacitation applied at the physical event seemingly led to further disruption rather than making the situation more peaceful (Waddington, 2011a & 2011b).

4) **Strategic Facilitation, also known as 'Dialogue Policing'**
   - Waddington (2011) presents an account of how the police liaise with different organisations during protests historically and how using a permissive approach to policing public order involves a greater emphasis on facilitating the right to protest. Currently, UK police use a 'Dialogue Policing' approach that looks to facilitate a protest and build relationships with community and protest groups to maintain the peace. Stott et al. (2013) suggest that the UK police have taken a step towards Strategic Incapacitation, but in a technical sense where the increased use of technology is being used to monitor an event.

Strategic Facilitation, as an approach to public (dis)order, has to a degree been a successful number of protests using this model have maintained order, leading to an increase in public safety (Stott et al., 2014). Stott et al. (2014) suggests this could result from the successful integration of Police Liaison Team (PLT) into pre-existing command structure and their operations. For example, the training of PLTs in negotiation and the public order-trained commanders has helped both to understand each other's defined role. In addition, the PLT's ability to build relationships with the protesters has helped to offer an improved quality of information on the ground to commanders, which would not have been reached by other means. PLTs appear to have only been tested on a relatively small scale of less than of never more than 5,000 protesters at an event, which has covered small geographical areas with smaller number of officers being deployed who were visible and accessible throughout (Stott et al., 2014). Therefore, questions about the effectiveness of the model remained unanswered such as, for instance, whether the skills of PLTs can be extended to other public order events, such as football and rugby. Additionally, how would LEA adapt their approach if there are larger intensified demonstrations consisting of tens of thousands of protesters. If there were a larger scale protest, it would be harder to circulate between members of the public and relate to larger numbers and, therefore, raised concerns about the safety of officers (Stott et al., 2014). As a result, the practicability of PLTs deployment at transient large-scale demonstration events must be questioned even though it has shown a good progress (Stott et al., 2014).

The psychology of police behaviour may be affected by training and influenced by their own psychological condition outside of work (Edwards & Kotera, 2020). There are many psychological factors that could affect a police officer's decision-making, such as personal bias and health, for instance, an officer may improve decision-making after eating lunch (Edwards & Kotera, 2020). In terms of actions utilised in demonstrations, the PLTs are trained to be negotiators to build relationships with the public in a particular way, whereas some officers are trained with riot-control tactics, such as

trudge and wedge (synchronised movements are performed to push against a resisting crowd) with the use of riot gear (Body-Gendrot, 2014). These different types of police behaviour and actions can impact a public order event, which unintendedly 'spark' disorder or contribute to it rather than maintaining the peace. For example, the reputation of PLTs should be maintained at a high level in order to keep that trust with the public in relation to their role; otherwise, if for some reason, the public feel that PLTs are deceitful based on their actions, then this could have a detrimental effect on how PLTs operate in the future.

Evaluating the Public Order Policing (POP) may be a technical matter dependent on philosophical orientation. The philosophy of POP is based on the mission and objectives of policing within the social and political context of an event (Wakefield & Fleming, 2009). In many cases it always depends on the relationship between the police and the people being policed (Wakefield & Fleming, 2009). The police operation during riots is important to analyse and to subsequently review strategical thinking; past riots in Burnley (2001) and London (2011) and the Sheffield Lib-Dem protest (2011) have been investigated to produce vital information. It is important that lessons are learnt from past mistakes and act upon them to improve the policing of public order, as *"effective (that is, trouble-free) public order policing is beneficial both to the police and the wider society"* (Waddington, 2007, pg8).

2.2.3 Model Consideration: analysis of public (dis)order events

Della Porta (1995) explains that it may not be possible to understand a protest's behaviour and its evolution unless there is an understanding of the context and interactions between police and the protesters. Therefore, analysis and evaluation of these public order events is important to determine more accurately how the police can adapt their service to the protest. The two main models that are used to analyse the phenomenon of public order events are the Elaborated Social Identity Model (ESIM) and the Flashpoint Model. Within the theory of both models is a central understanding of specific contexts and their dynamics (Body-Gendrot, 2014). In addition, these models try to address the root causes of social deprivation or political reasons of protests to identify the dynamics which have 'sparked' disorder and what has caused it to sustain the level of disorder, such as mobilisation of resources to cause greater violence.

These two models will be explained and evaluated to illuminate how they are used to understand and analyse public (dis)order events: -

- Elaborated Social Identity Model (ESIM): The ESIM is a sociology theory to explain crowd behaviour based around group interactions, as collective action is likely to happen when individuals share a social identity (Challenger et al., 2009; Wijermans, 2011). The ESIM is currently used in policing of public order practice, as its simple to use and, importantly, helps the police to identify main

processes involved in the emergence and escalation of collective conflict in order to provide a framework to reduce the probability of disorder (Challenger et al., 2009; Wijermans, 2011).

- Flashpoint Model: The Flashpoint Model has seven levels of analysis (structural, political/ideological, institutional/organisational, cultural, contextual, situational, and interactional) to describe the relationship between disorder and the narrowest of interactions where flashpoints commonly occur (Waddington et al., 1989; Waddington, 2007). The Flashpoint Model attempts to integrate levels of analysis that are used to identify why some disorderly flashpoints fail to ignite, while trying to elucidate why similar or different incidents trigger a 'spark' leading to a public disorder event (Waddington et al., 1989; Waddington, 2007). To understand "why disorder does or does not occur" is a means to use an interpretive framework (Jordan, 2016).

In comparison, neither ESIM nor the Flashpoint Model analyse at an individual level (Wijermans, 2011). Where ESIM analyses crowd behaviour focus on group interactions that share a social identity, the Flashpoint Model has seven distinguished levels that identify various complex phenomena dependant on the situation (Wijermans, 2011). The Flashpoint Model is seemingly not used by the police as its more complex to understand (Wijermans, 2011). its complexities are less easy to interpret and are difficult to put into practice within a fast-changing sequence of time-orientated events on the ground. The ESIM model is more widely accessible and easier to understand due to its simplistic approach in how the model is clearly outlined to its audience, where the transformation between theory and practice seemingly makes it more effective to act upon (Wijermans, 2011). The Flashpoint Model can be used retrospectively to analyse events before, during and after one has taken place to identify the 'spark', but its model could be used to predict capability to see what may happen next in real-time events (Jordan, 2016).

Both models appear to encounter difficulties in application, as some demonstrators/ protesters/ may question the legitimacy of police involvement (Wijermans, 2011). This makes it more difficult to control the crowd effectively to prevent public disorder. There is a much deeper social problem that has to be worked upon to restore trust in police, otherwise these tactics used may be less effective. Restoring trust in police may make it easier for effective management of public order events to increase public safety (Wijermans, 2011).

2.3 Police and Social Media

2.3.1 Social Media

Social media comprise highly interactive websites and applications that can be consumed by different ubiquitous technologies. This enables more individuals and communities to participate in creating, sharing, discussing and modifying user-generated content and being informed (Ji, 2010 & Kietzmann, 2011). As a result, regardless of geographic location, social media continues to grow (Clement, 2020; Ji, 2010; Kietzmann, 2011). The speed of responsiveness of the users expressing political, economic or religious views on Twitter or Facebook, where they interact post and share information has led to these becoming viewed by some as a valuable source of public opinion. These existing social media networks, such as Twitter, Facebook and Instagram have a rich and diverse ecology that have a different scope and set of functions (Georgakopoulou, Iversen, Stage, 2020; Kietzmann, 2011). Facebook, for example, seems to appeal to those users who look to connect and share with people in your life in a more informal way (Borgatti, Everett, Johnson, 2018; Kietzmann, 2011), whereas LinkedIn is viewed as a network to connect *"the world's professionals to make them more productive and successful."* (LinkedIn, 2015). The growth of social media has a significant impact on organisations' operations, reputations and both positive and negative relationships, which shows social media has brought new challenges (Georgakopoulou, Iversen, Stage, 2020; Kietzmann, 2011). Organisations, such as the police force, may need more time to thoroughly understand social media different forms and possess the skills to engage effectively on social media (Oscar, 2018).

2.3.2 Police use of Social Media

Social media has been used within past years in LEAs, and transforming performative operations, relational engagement and interaction with the public (Akhgar & Staniforth, 2014). LEA usage and understanding of social media has grown in countries as shown in the UK, United States of America and Netherlands since the rise of social media. For example, GMP's social media approach is developed to a greater extent and is overall more effective than other agencies such as the MET (Crump, 2011; LexisNexis and IPCC and Bartlett et al., 2013). Despite, this really good use of social media by some forces, a recent report produced by the Open Source Communications Analytics Research Centre (Oscar) (Oscar, 2018) outlined *"approaches to social media were fragmented and some forces struggled to keep up with technological advances"*. Martin Innes (Professor of Police Science) has also added that police services are struggling to *"keep up with the changes and disruptions that are being caused by social media"* and that the police need to rethink how they approach the Information Age (BBC, 2017; House of Commons, 2018). He (BBC, 2017) noted it was surprising that this was not a mainstream position considering the positive examples set out by specific

police forces using social media to build trust, engagement and rapport with the public as a way to improve crime detection rates. Within other countries, notably Queensland Police in Australia and New York Police Department in the USA have adopted humour into their social media approach to improve relations with the public to gain trust and improve levels of engagement and reputation (IACP, 2014 & LexisNexis, 2012 & 2014). The social media phenomenon has significantly impacted both positively and negatively upon forces operations, reputation and relationships (Kietzmann, 2011; House of Commons, 2018).

The LEAs have made demonstrable progress in adoption of social media to support police activities, such as aiding in criminal investigations (IACP, 2014; LexisNexis, 2012 & 2014). Although these countries LEAs (e.g. UK, USA, Netherlands and Australia) have made progress with the effective use of social media, priority seems to be given to crime prevention and investigation rather than to in-service training and listening/monitoring for criminal activity (HMIC, 2011b; IACP, 2014 & LexisNexis, 2012 & 2014). It would appear that LEAs require greater effective participation, resourcefulness, understanding and utilisation in their approach to social media (HMIC, 2011a & 2011b; IACP, 2014; LexisNexis, 2014a; Metropolitan Police Service, 2012).

2.3.2.1 Information Sharing

Traditional police-to-citizen communication utilises channels such as news media, leaflets, face-to-face interactions and meetings to disseminate their message, but now social media has permitted more instantaneous transfer of information. Social media enables the police to share information in real-time where their audience can read and share posts, which means the cost of dissemination is low. In order to disseminate and gather information, the police must Increase public awareness to gain followers or 'likes' to extend impacts on the public. These posts can inform readers of criminal activity, provide reassurance by refuting rumours and provide regularly update briefings on an unfolding incident to proffer safety advice to the public to minimise risk and improve areas of safety (Crump, 2011).

2.3.2.2 Engagement

Preliminary research shows that several police forces, notably Greater Manchester Police (GMP) and West Midlands Police (WMP) in the UK have been using and experimenting social media since 2008. Initially This was started as an initiative started by officers, with some official support, but over time additional support came from the Association of Chief Police Officers (ACPO) (Crump, 2011). UK LEAs recognised that social media need to be taken more seriously after two major events occurred, the Tuition March in 2010 and the widespread riots in 2011 (Crump, 2011; Downes, 2013). Since these demonstrations, police usage of social media has increased and extended

across forces/departments; according to Downes (2013), most British police forces operated an official co-operate account, with 98% having a presence on Twitter, 96% on Facebook, and 94% on YouTube, with others assessing platforms. WMP's data analytics, for example, highlighted that 13–17-year-olds comprised the smallest percentage of followers so to improve interactions with this group now use SnapChat which is more popular with younger users to improve relationships and trust between communities and the police (Eccleston, 2016). The GMP and MET forces use social media as an opportunity to share information, dispel rumours, and to reach many other groups to crowdsource information to help answer police enquiries, such as with the outbreak of looting in London and Manchester 2011 riots (Crump, 2011).
The police's communication with the public is known to be important, as the public's co-operation can help the police effectively maintain order and reduce crime levels (Crump, 2011; Torre et al., 2018). Social media aids the police to increase engagement, transparency and legitimacy, collaboration, community participation, reputation, and to communicate and interact directly with citizens by posting interactive content, such as videos and images (Torre et al., 2018). This may help encourage the public and other organisations to positively help police their community through vital information. It is believed this may allow for greater trust and confidence in the police to aid in a higher level of interaction at offline and online at public order events (Torre et al., 2018). The police's individual and organisation accounts can use social media to reach out directly to their community and also further afield to communicate with demographics in differing locations. The public come to recognise the faces of officers and to personally relate to their local police force, increasing both relevance for community policing and levels of engagement. This requires greater resource, but as a result of police cutbacks it appears to be difficult to enforce a larger online presence to provide higher level of availability of crime statistics to increase engagement, public accountability of policing and address public misconceptions of crime pervasiveness (Longstaff et al., 2015).

The importance of online presence is evident as a younger audience has expressed greater interest in contacting police online (London Assembly, 2013). Social media may facilitate an increased connection to the public, as it may encourage the disinterested to participate in policing their community (Accenture, 2012). A survey by Accenture (2012) indicated that 69% of UK respondents said they would more likely contact the police via social media, as they could remain anonymous. As a result, the police may receive information for their investigation, which possibly not have acquired before without this medium (Houses of Commons, 2018). The Accenture (2012) survey also suggested that 58% of respondents would like a police presence on social media to provide information and also to engage more frequently with their community, such as uploading photographs of criminal activity and providing progress checks on particular relevant investigations (Houses of Commons, 2018). Bullock (2018) argues that the police *"have not yet served to facilitate interaction between constabularies and citizens in the ways that have been proposed and desired, the*

article considers factors that structure the transformative potential of social media." Additionally, Bullock notes that the process to implement social media practice is challenging and emphasised that a focus on technology itself will not bring about organisational change for the changes required to transform police and citizen engagement.

Police organisations may find it challenging when developing a strategy for social media, as its can be difficult to know how to handle and control general communications and two-way interactions, and what appropriate communication method the police should adopt when on social media (Crump, 2011; Fernandez, Dickinson, Alani, 2017). An online presence that interacts can help form a personal connection with the public and facilitate a positive attitude similar to a PLT's role within the Strategic Facilitation Model (Longstaff et al., 2015). The level of interaction between the police and public on social media is of a low frequency due to cutbacks within the police and resources being limited (Pickles, 2015; IPC, 2013). This issue is not limited to the UK, for instance, South Wales Police (based in Australia) allowed their citizens to report crimes, raise concerns and hold virtual meetings via an app on their device. However, the issue with this project was the limited engagement of police officers' responding to citizens in a timely manner. This situation highlights how that resourcing issues can have an impact on the police's ability to engage with the public effectively on a wide scale (Pickles, 2015; IPC, 2013). There is a risk in the police offering an interactive service, as it can raise expectations from citizens to provide a higher level of engagement that cannot be met. Bartlett et al. (2013) have found the police emphasise citizens should not report crimes on social media and revert to 101 or 999. The research suggests that tweets do not receive the same level of urgency as other communication streams and due to the fact that they do not integrate their social media accounts into the force's control centre.

The Accenture (2012) survey highlighted that a fifth of their respondents were aware of police presence on the social web. Public awareness of the police's use of social media and digital technologies has to increase in order to extend that level of public participation and engagement online in dialogue about local policing (Accenture, 2012; House of Commons, 2018; IPC, 2013). As a means to deepen and expand local democracy, technology will not answer everything, as good experience with news media to strengthen citizen's awareness and participation in public services is on the rise. This should be incorporated into police practice making them locally responsive and accountable (Accenture, 2012; House of Commons, 2018; IPC, 2013).

2.3.2.3 Intelligence

Social media and the changing nature of new technologies are impacting on police operations. The audience using these social applications are growing exponentially alongside their fast responsiveness in situations is challenging the management of

public disorder (Baker, 2012; HMIC, 2011a & 2011b; House of Commons, 2018; Murji, 2011). Similarly, HMIC (2011b) concludes that there is a *"strong possibility the current focus is inducing a situation where police are "unsighted" in respect of a range of risks."* (HMIC, 2011b, p38). This quotation seems to suggest that the police require a greater technological capability to identify wider intelligence to increase their risk awareness to potentially prevent crime (HMIC, 2011b; House of Commons, 2018).

Torre et al. (2018) outline that social media is used to prevent *"unrest and signalling suspicious situations (intelligence), and as a tool to collect and analyse large quantity of open data to improve crime and prevention"* The idea is for LEAs to use social media analysis as a tactic to prevent crime, but one must be careful for the method not to become *"undemocratic or unauthorised surveillance of citizens"* (Torre et al., 2018, p.g.6). This process of gathering intelligence and knowing whether it is credible is complex and uncertain due to the nature of Big Data and other issues, such as the number of fake profiles created on social media that seemingly spread false information (HMIC, 2011a & 2011b; House of Commons, 2011 & 2018; Hurwitz, 2013; Innes & Roberts, 2011, Longstaff, 2014). Additionally, the police face some difficulties working alongside social media organisations, for instance, if police request information on immediate threat to life. This will be immediately shared, but social media platforms do not judge the credibility of what's been said on social media. There is no credibility threshold when measuring a situation, so no notifications are required to inform LEAs apart from when information is requested; otherwise for any other the police request to access registered user's data could be problematic and time consuming to process it due to certain organisational and jurisdictional laws (Parliament, 2016). In addition, materials posted via social media platforms of a criminal nature may be removed by social media platform algorithm automatically and not necessarily notify the police (Dencik et al., 2015). If this happens then the police might not be able to prevent, for example, an individual being murdered. Partnerships with social media organisations have been created, such as the police partnering with Facebook in the fight against child abuse and child pornography by prevention and prosecution of offenders (Longstaff, 2014).

Partnership between police and other law enforcement agencies is a key element to help prevent crimes, so, therefore, improved engagement is required between each organisation for this to happen. These social media platforms are private and run on a profit basis and have a duty to protect their customers according to policy and procedures. Since 2013 when Edward Snowden revealed secrets about private organisations creating backdoors for law enforcement agencies, the public feel their privacy is being violated (Privacy International & Amnesty International, 2015). The balance between privacy and security is a key area of debate for organisations, such as Facebook and Apple reforms making it more difficult for various LEAs to access potentially vital data for progressing investigations (Privacy International & Amnesty International, 2015). However, when law enforcement accesses this data there is a

need to identify and understand the moral and legal obligations to retain and disclose data (Babuta, 2017; Torre et al., 2018). In addition to this, the police need to have a greater understanding on who holds the data to improve in accessing it in time. A solution may be for an international liaison between governments to decide on what information can be disclosed and form an information sharing agreement, so the LEA's investigation can proceed speedily to prevent crime. The public seem to expect the police to act and investigate this evidence from a social media platform. The police and Government agencies need to improve educating the public on how an actual investigation process is handled and what the public can to help in this process (N8 Policing Research Partnership, 2015). Education can help build trust in what the police are doing rather than being against the system, potentially making it more difficult for the police to conduct their job to prevent crime (N8 Policing Research Partnership, 2015). Even though the police improve their engagement and the public gain more awareness, social media platforms, such as Facebook and Twitter need to continually improve their platform to increase public safety (N8 Policing Research Partnership, 2015).

A high number of events are seemly monitored by the police on social media (Dencik et al., 2015. This can involve the collection of social media data leading up to, during and after an event. Social media monitoring can be used for pre-emptive measures and real-time police tactical and operational responses (Dencik et al., 2015). The police seem to monitor events with prior information of an event occurring through various forms of intelligence, media and knowledge of community tension. Currently, the police refer to a "disorder model" to explain the nature of disorder, which may aide in the management of policing operations, events and incidents where the risk of disorder or potential for disorder to escalate (College of Policing, 2018).

Tension or a level of disorder might be present in any community and social grouping. Its management is a continuous partnership *"rather than one of crisis intervention involving the police as a single enforcement agency"* (College of Policing, 2018). According to the ICOCO (2011) report, it seems that the level of social media analysis does not include general monitoring of community activities. This extra layer of detail might help enhance police engagement and assess the needs of the local community. This layer may have been already implemented as the paper cited is six years old. However, three years later after the publication of the ICOCO, Dencick et al. (2015) suggested that the police are still assessing whether to employ *"social media monitoring for potential tension surrounding the police, or hostile mentions of the police, what was described as 'looking for reputational risk for the force."*

Intelligence-led policing can affect the police's priorities and tactical decision-making when adapting to a public order event. Social Media Intelligence (SOCMINT) is a form of new intelligence to ensure public safety, but it raises ethical, operational and technological challenges for LEAs (Bekkers et al., 2013; Miller et al., 2013). For

example, SOCMINT abides by a legal regulation, such as The Regulation of Investigatory Powers Act 2000 (RIPA), when data is being collected and analysed (Miller et al., 2013) there must be an *"effective and necessary contribution toward security and safety"* (Omand, Bartlett & Miller, 2012, p.g.8) that is *"proportionately and appropriately balanced against other desirable public goods – such as the right to private life."* (Omand, Bartlett & Miller, 2012, p.g.8) Importantly, these laws that are applied to protect society could be based on a *"form of public acceptability and involvement"* (Omand, Bartlett & Miller, 2012, p.g.8). As relevant regulations are constantly updated, the police must update their understanding frequently to ensure that they remain in compliance with the law when conducting overt surveillance/ monitoring of the public domain sources and protecting the public's anonymity in police operations (Miller et al., 2013; Omand, Bartlett & Miller, 2012). Despite this issue, social media can offer a way to crowdsource intelligence, providing more content as information is created and shared on a large scale (Miller et al., 2013).

The intelligence gathered from social media sites, and the speed of posts/messages can be vital when appealing for information regarding a missing person or other criminal activities to assist in the police's investigations. The monitoring of social media can provide insight into police forces and their communities, gather data on suspects from their individual social media accounts giving further insight on their location or circles of friends and identify potential issues that assist in polices investigations (HMIC, 2011a & 2011b; House of Commons, 2014b; McCarthy, 2014; Innes & Roberts, 2011). Social media may assist LEAs to trace individuals involved in incidents by using systems such as Facebook, Twitter, Flickr and Instagram, especially when citizens and the media are witness to these events and are documenting them through video and photography. SOCMINT can help to improve the quality of the police's intelligence, thus potentially optimising effectiveness and make timely decisions when deploying their tactics to a situation (HMIC, 2011a & 2011b; House of Commons, 2014b; McCarthy, 2014; Innes & Roberts, 2011). A review of relevant (academic/research) literature suggests the LEA need to improve understanding of social media and the adoption of formal policies and processes to enable a unified and consistent approach to using modern technology (Torre et al., 2018). Thus, with such improvements, police forces might be able to adapt faster to the changing situational complexities of real-time events, so less or no disruption is caused (HMIC, 2011a & 2011b; McCarthy, 2014; Innes & Roberts, 2011).

2.3.2.4 Enforcement

As previously outlined in section 2.3.2, social media is used by the LEA to track down offenders. Social media is considered a source of information when criminals leave evidence using it. It is known that police forces internationally monitor the dark web websites promoting hate and forums to ferment anti-social behaviour to gather evidence for investigations (Torre et al., 2018). The population on social media has

grown and people are spending more time on these social platforms than any other form of activity online (GWI, 2014 & 2015; Miller et al., 2013; SkyNews, 2015). SkyNews' (2015) freedom of information request to police forces showed an increase in social media crime. The analysis of this data showed that threats to kill, harassment and sexual offences have seen a significant rise on different social media platforms between 2012 and 2014. For example, GMP reported 495 crimes on Facebook in 2012, and saw an approximate double increase in 2014 as 959 crimes were reported (Birchley, 2015; SkyNews, 2015). It may be suggested that police enforcement on social media may not be effective enough. Moreover, the police are constantly under pressure to keep up-to-date with revised and new laws both on-line and off-line. This can make police investigations challenging when trying to maintain consistency when approaching people's privacy (Bekkers et al., 2013; Miller et al., 2013; Omand et al., 2012). There are further challenges presented to enforcement on social media in that there that are resource limitations, jurisdiction issues and different governance/compliance issues regarding staff management.

Police often are exposed to sensitive information that cannot be divulged. Members of the police within different positions, levels of rank and experience who use social media must be aware of the impression being projected as a representative of the police (Police Foundation, 2014). Furthermore, the police must ensure that appropriate content is placed into the public arena (Police Foundation, 2014). Each of the UK police forces has documentation that appears to vary in providing their own different social media strategy rather than conforming to a unified national approach (Police Foundation, 2014). This may cause conflict and confusion when police are using social media. The Twitter accounts of a number of police officers in Northamptonshire were banned as tweets contained information that breached the Data Protection Act 1998. The information tweeted may have harmed both a police investigation and violated the law (Scott, 2012).

The HMIC (2012) report identified 357 possible instances of inappropriate police behaviour conducted on social media organisations within a nine-month period, in which 71% occurred on Twitter. The type of behaviour and number of instances related to: 1) Offensive language/ behaviour: 132 2) Comments against police procedure: 119 3) Negative attitude towards work: 70 4) Extreme opinions referring to the government: 39. The HMIC report further outlined that nine police forces had the capacity to monitor personal staff accounts, and nine failed to make any checks. At that time, HMIC (2012) recommended that police forces should apply appropriate monitoring, managing and training of police officers to use social media. In addition, the former lead on digital engagement, Deputy Chief Constable Gordon Scobbie emphasised that police officers should be trusted to use their social media account and given training and support to ensure appropriate interaction with the public. If police mistakes are made, then suitable allowances should be provided use of social media is a learning curve, without failure the police will not develop and succeed (Laville, 2012).

National Policing Lead on Digital Engagement, Deputy Chief Constable Ian Hopkins recognises social media is an integral part of modern society that can be used productively to influence crime reduction and detection, as its speed and reach is powerful, but the posts on social media may not represent the wider view, but it can encourage discussion, debate and give insight (NPCC, 2014). A unified set of rules and procedures may be required to govern all police forces in the UK to ensure conformity to reduce confusion surrounding acceptability, leading to minimise mistakes.

Despite a number of challenges regarding enforcement, social media does introduce a new source of evidence for enforcement and the justice system to prosecute criminals (LexisNexis, 2014b; Miller et al., 2013). For example, Michael Grasso, a Sicilian drug dealer, had evaded police capture since 2010, but was arrested and deported a few years later due to social media activity, which included a photo, name and location for where Grasso was known to be working (Miller et al., 2013). Similarly, MPS gained crucial support from Twitter during investigations in the 2011 August London riots, as Flickr images and videos of potential suspects (uploaded by police) had been re-tweeted 8,500 times and viewed on 4.3 million occasions. As a result, large numbers of the public proffered information and suspects even turned themselves into the police. In time, the police were able to build up a key list of community contacts and empower the local people to cooperate in problem solving in order to maintain control keeping the peace within the community. In conclusion, it can be seen that social media has its positive and negative aspects that the police must respond to lawfully prevent social media crime occurrences (Miller et al., 2013).

2.3.3 Social Media Audience

It is important to detail elements of the different user groups when using social media as an information source for intelligence services. Findings from a number of recognised organisations will be presented with some based within the UK such as weareFLINT, Office of National Statistics (ONS) and Statista, 'Pew' based within the USA and the Global Web Index (GWI) covers a wider set of countries. [1] Each of these platforms releases organisational information, describes its user characteristics and highlights the most actively used platforms and largest number of registered users. However, these platforms have been chosen on degree of relevance to the LEA as outlined in section 2.3. The significant impacts of these platforms upon the public will be presented to further evidence the reasoning behind the focus of LEAs attention being primarily on these platforms: -

- Facebook (Over 1.94 billion users (CNN, 2017))
    - 32 million users are in the UK (Social Media Ltd, 2016) and has 78% of accounts are actively in use according to weareFLINT (2016).

[1] Pew is it not an acronym, but USA based statistics

    - Facebook is used more often by women than men (Pew, 2016), but according to ONS (2016) and weareFLINT (2016) shows similar levels of usage across age groups. In 2016 the ONS (2016) shows highest number of users fall in the age group 25 to 44 with the fewest number 65+ users, but has shown an increase. According to Pew (2016) and weareFLINT (2016), the usage can vary between groups. For instance, if an adult has both higher educational levels and household income, and lives in an urban area then they tend to show a greater usage of Facebook. However, there is a couple of percentage difference between location areas and a slightly wider gap for the other categories.
- YouTube has approximately over 1 billion internationally registered users
    - YouTube has around 19,100,000 million users in the UK (Social Media Ltd, 2016) and 85% of these are actively using the platform, according to weareFLINT (2016).
    - YouTube reaches more 18 to 34 and 18 to 49-year-olds than any of the cable networks in the US (YouTube, 2017; weareFLINT, 2016). According to the ONS (2016), it shows a similar trend to Facebook in terms of the age groups in the UK.
    - YouTube users in living in an urban area (not much difference of usage on percentage based on location) and have a slightly higher income show a greater usage.

- Twitter does not open disclose the total number of registered users, but states it has 328 million active users (Twitter, 2017b; Statista, 2017).
    - 15 million users are based in the UK (Social Media Ltd, 2016) and 45% are actively using Twitter according to weareFLINT (2016).
    - Twitter is frequented mostly by younger adults (ONS, 2016), and usage increases if the user lives in an urban area and is more highly educated, but there is near equal usage in terms of gender (ONS, 2016; Pew, 2016).

According to GWI (2016), digital consumers spend on average 1hr and 58mins per day on social networks and messaging. This level of daily media activities results in networking/messaging consuming near 1 in every 3 minutes spent online. The average length of time spent on social media is increasing and has seen a rise of 20 minutes since 2012. ONS (2016) suggests that the mobile devices most used to access the internet in the UK are a smartphone (70% share), followed by portable computer, (36%) and other handheld devices (12%). Furthermore, in Figure 2.2 two thirds of the majority use social media at least once or several times a day.

[Intentionally Left Blank]

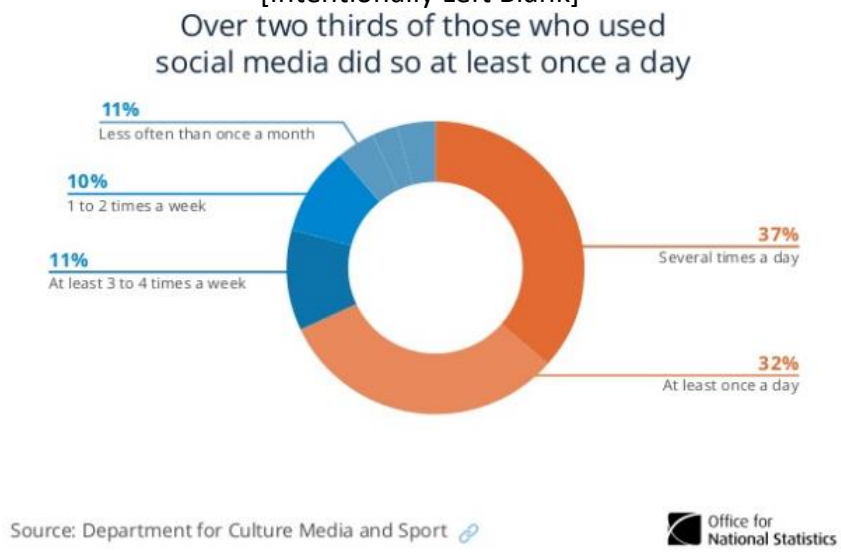

Figure 2.2 Number of times users actively use social media (ONS, 2016)

Specifically, in the UK, there has been a significant 22% increase in Internet use of social media from 45% in 2011 to 63% in 2016 (ONS, 2016). The disparity in the usage figures based on age, gender and education levels emphasises that social media platforms differ in the users they attract. Such demographic features affect frequency of posts, type of content and its style. ONS (2016) suggested the approach of the end-user can be different towards different social networks.

2.3.3.1 Rationales for Using Social Media

The main purposes of online activities are socialising, but some users online seek anonymity. Social media is used not in a sense to create new relationships, but instead maintaining existing ones (Campbell and Kwak, 2011). It has been estimated that between 85-98% of participants use social media to maintain and re-enforce relationships that exist offline, comprising family, friends or people with similar interests (Lenhart et al.., 2013). ONS (2016) suggests the most common reasons for individuals in the UK to use social media are to find out what happens in the local area (41%), share content and related views (29%), meet people (17%), locate to do activities (17%) and to chat about interests (16%), such as sports/ art/ music. According to Brandtzaeg (2012), individuals can differ variably in their approach when using social media and defines five distinct user types:

1) Sporadic users: This group of people are low level users, who rarely connect with a Social Media Platform (SMP), but may use it to check if a family or friend has been in contact.

2) Lurkers: This group are individuals who use SMP(s) to passively consume content of others by, for instance, to view photos, find information about

friends, see if somebody has contacted them or simply to fill time. This type of user does not tend to interact or contribute.

3) Socializers: Such individuals are primarily interested in using SMP(s) for social interaction with family and friends.

4) Debaters: These individuals mainly use SMP(s) to participate in debates and discussions by actively writing and uploading personal contributions.

5) Advanced users: These individuals use SMP(s) on a frequent basis for multiple purposes, such as contributing, debating and socialising, and demonstrate the broadest and most varied array of behaviours.

These user types are useful when considering variations in engagement with a series of actives put forward by the LEAs, for example, requesting help in investigations. The means by which the police engage on social media may require different approaches to attract say socialisers compared different user types. The subtlety in differences may be related to socio-demographic background, as it is reported that individual with higher levels of education are the most active user groups. It may be the case that social media content may be biased towards these groups, whereas the disadvantaged groups might be difficult to reach.

2.3.3.2 Role of Social Media, Organisations and Activism

Amin (2003), Eadson (2011), Greer (2010), King and Waddington (2004), Institute of Community Cohesion (2010), Waddington (2012) researchers have established variables may cause a protest or demonstration, such as the economy, unemployment and social depravation. As a result, the dimensionality of variables in an event between the police and protesters can present situational complexities that challenge the management of public disorder (McCarthy, 2014). The delicate nature of the protest or demonstration may be more vulnerable and susceptible to an influence that could trigger a volatile reaction to cause sufficient increase in tension to lead to a degree of rioting (Amin, 2003; King & Waddington, 2004; Greer, 2010; Rosie, 2009). For example, this adverse reaction may have been caused by further agitation of inflamed comments, inadequate police tactics, police perceptions and indifferences, news coverage and press sensitisation (Amin, 2003; King & Waddington, 2004; Greer, 2010; Rosie, 2009). The unprecedented speed of social media alongside the widespread and ever-increasing use of mobile devices has distinguished recent demonstrations and riots from previous disturbances. There have been significant concerns about the role of social platforms, such as Twitter, Facebook and YouTube in relation to the organisation and spread of the disorder (Lewis et al., 2011). Importantly, Baker (2012) suggests that social media does not cause people to riot:

"Yet, while social media can help to explain the speed and the capacity to orchestrate riots in many cities across England, social media cannot account for the failure of attempts to organise riots via social networking sites in areas, such as, Plymouth in southwest England, and Northwich in northwest England. Or, as said another way, while new social media contributed to the form and effect of the riots, they were not the initial cause of the civic unrest. Social media/ networks *"does not cause one to riot, just as being a member of Twitter or Facebook does not make one more susceptible to violence." (Baker 2012, p.45)*

Although social media may not always cause hindrance to public (dis)order events, there are reported incidents in which the public are positive in their support when witness to incitement via social media (Baker, 2012). Incitement through social media may not cause change to people's behaviour to commit an act for an event. In this case, many people make use of Twitter to discuss and disseminate events through retweets, which include tweets from official sources such as the mainstream media (Baker, 2012). Social media has been used in the aftermath of an incident for the greater good of the community and country (Baker, 2012).

Activism is defined as *"the activity of working to achieve political or social change, especially as a member of an organization with particular aims" (Oxford Dictionaries, 2020),* two regularly encountered types of activism are: -

- **Physical Activism:** Physical activism utilises traditional methods, such as face-to-face meetings, radio and television to start a citizen movement toward a goal, objective or cause (Treré, 2018).

- **Digital Activism:** Digital activism is observed *"where digital tools (the internet, mobile phones, social media etc) are used towards bringing about social and/or political change."* (Rees, 2015). Digital activism is also known as 'Cyber Activism'. According to online activism think tank, Meta-Activism Project, digital activism seems to serve six key functions, which are: *"shaping public opinion; planning an action; sharing a call to action; taking action digitally; transfer of resources."* (Rees, 2015).

Social media platforms exist in different forms and some notably activists, have been using tools for change, collaboration, co-ordination, information-sharing and protest (Rees, 2020; Sivitanides, Shah & Marcos, 2011; Treré 2018). Protests still take place in physical public spaces, but social media is a tool that can be used to increase publicity of a cause, share information and updates from meetings to current and potentially new supporters (Rees, 2020; Sivitanides, Shah & Marcos, 2011; Treré 2018). Popular hashtags are used in protest events to help centralise key information, identify prominent activists and journalists as sources for updates and analysis (Treré 2018). Physical protesting is a key part in the success of social movement for trying to cause

political change, but social media also provides an arena, as it may help provide global attention to a cause (Rees, 2020; Sivitanides, Shah & Marcos, 2011; Treré 2018). This may vary from information provided by local media sources and from that desired by current political agenda.

The influence of social media is apparent within the news media, such organisations regularly track the conversational developments that draw attention on social media, and consider influences on offline activities (Rees, 2020; Sivitanides, Shah & Marcos, 2011; Treré 2018). Online action through clicktivism may cause digital distraction, which the participation online could potentially halt offline action, but may have the ability to increase mobilisation for the event (Rees, 2020; Treré 2018). There are issues in turning online support to action offline, such as how would a 'like' on Twitter or Facebook be turned to offline protest behaviour (Rees, 2020; Treré 2018). There are other issues of concern, such as the hacking well-respected social media accounts for duration of an event in order to influence it (Rees, 2020; Treré 2018). It is important to establish in such cases who the perpetrators are in that timeframe; otherwise it could have negative effect. In addition, it would be useful to assess whether offline activity of individuals exerts similar levels of influence online (Phillips, 2016).

The speed of communication on the social web may result in the participation of larger numbers in protests and the subsequent spread of protests or riots elsewhere (Greer, 2010 & Rosie, 2009). For example, a protester may be cornered through the use of "kettling" by the police for hours without reason, leading to this individual to use their mobile device to post a negative response on social media. This post may cause an emotive reaction both on-line and off-line possibly creating further tension and disorder, as other protesters may respond to this message positively or negatively. What happens on-line is important as to how people react off-line, as either side may be contagious and lead to indignation being infectious (Greer, 2010 & Rosie, 2009). The interaction of livestreaming of content on-line with offline actions relies on online co-ordination and communication in an event (Treré 2018). These online spaces can connect and maintain links with other geographic locations in which there are supporters for a particular cause that relate with their social movement. For example, in the USA, Baltimore's #blacklivesmatter campaign spread to other locations, such as Ferguson, Philadelphia and New York, as a result of the use of social media and news outlets. The constant on-line presence of the group facilitated the growth of the movement (Choudhury et al., 2016).

Social media can allow the fast exchange of communication and information that has not been evidenced in previous historical cases. This has caused commentators to blame social media for widespread disorder (Baker, 2012). However, Baker (2012, p.g.45) argues that *"to blame technology as the cause of the riots is accordingly limited. Riots have occurred at regular intervals in modern Britain long before these technological innovations, and while new social media facilitates social networking in*

diverse temporal and spatial boundaries, it is a facilitator rather than the underlying cause of collective action". Technology enables information sharing but, ultimately, a riot is conducive to the topic of debate (Baker, 2012; Treré 2018).

Prior to the use of mobile technology, public gatherings enabled people to voice opinions, discuss and disseminate a particular message. Social media may not cause a riot, but it can provide a means for instantaneous connection for people on a large scale, and an arena in which to observe physical riots via citizen journalism. Citizen journalism aims to produce information that challenges the 'official' version of events (Greer & Mclauglin, 2010). The inclination of citizen journalists and professionals is to actively find and collect, disseminate and analyse information communicated in the marketplace that commodifies and on mass consumes the adversarial news (Greer & Mclauglin, 2010). It may be argued that this helps build momentum and bring about social change and giving a voice to those that would usually have to struggle to be heard, but on-ground presence is seemingly needed, to give action to the words (Greer & Mclauglin, 2010).

Social media has changed how people are informed, work and communicate with one another at any time (Castells, 2009; Chui, 2012). This is where 'citizen journalists' have come forth with information that is collected, curated and published on social media and blog websites, that is shared and consumed by the public (Greer & Mclauglin, 2010; Treré 2018). Citizen-generated content (information and images) can provide for endless mashups, remixes and altering of events, where the news is redefined by the driving force by the rapid mobilisation of the citizen journalists. Feedback can be quickly provided by an individual on what the citizen journalist posted (Greer & Mclauglin, 2010; Treré 2018). This social change has had positive and negative effects on institutions and the public in society which has altered its stance in this era of social media. New forms of identities, cultures, language, affiliations and movements have spread online, and at times has led to being in the spotlight offline in society (Greer & Mclauglin, 2010; Treré 2018).

In past events, such as 2011 London riots, citizens who captured video footage of riots were encouraged to share it to the police to aid investigations. Facebook pages were created to identify looters, whilst websites such as Flickr were used by the police to find images of looters (Couts, 2011; Guardian, 2011). These rioters failed to cover their faces and even posed for pictures with stolen goods, posting them on SMP(s), thus making it easier for the police to investigate individual incidents (DeCastella & McLatchy 2011). In other circumstances, citizen journalism has helped to contradict police statements, which has caused many police forces to draw upon ACPO and NPIA guidance to improve engagement (Greer & Mclauglin, 2010; Treré 2018). Citizen journalists have a responsibility too, where the individuals must be held accountable for their own actions in-order to prevent rising tension and providing/ spreading mis-information that could harm the public (Greer & Mclauglin, 2010; Treré 2018). It can

be seen that caution should be taken to the content transmitted on social media, in distinguishing fact from fiction. Education is the key to help guide citizen journalists and to help bridge the gap between police and the citizen journalists to increase public safety (Greer & Mclauglin, 2010; Treré 2018). The UK police have had to modify their command-and-control operations due to speed of communication on social media, meaning the police have had to quicken their police response (HMIC 2011a & 2011b; Houses of Commons, 2018).

Journalism has a professional role to report the news of the events. This news can have a direct influence on the events that led up to the events, during and after they have unfolded (Greer & Mclauglin, 2010; Treré 2018). Journalists do use social media as a source to see what is happening within events through the discussion or postings on various social media platforms (Greer & Mclauglin, 2010; Treré 2018). While it is useful to make use of this media, but it can be a bad if stringent verification checks may have not happened and be used as an easy replacement for interviews. There are many people in positions of great power on Twitter, but their presence on social media can be difficult to verify and make them accountable for their actions (Greer & Mclauglin, 2010; Treré 2018). This can lead to negative consequences, as it means journalists cannot interact or hold discussions with those individuals. Those people in power on Twitter can, with only a few tweets, change a series of events leading to positive or negative events. These tweets can possibly satisfy some of the public and journalists. This may be alarming for journalists who might have covered these individuals for years but may have never interviewed them in person (Greer & Mclauglin, 2010; Treré 2018).

The Government has become more aware of the power of social media over traditional media for the better or worse. The posts and interactions on social media could be seen by anyone, which has led to some concerns about the use of surveillance of the movement and its participants. Research on the Occupy Wall Street demonstration in New York confirmed this thought, as activists showed concerns about government and police surveillance (Penney & Dadas, 2014). Individuals chose not to post about that particular event on social media because they felt these channels were being monitored (Croeser, & Highfield, 2014; Penney & Dadas, 2014). The concerns of social media surveillance, such as its public nature of tweets and real-name connections on Facebook, led to activists in censoring their posted opinions, information and links online (Croeser, & Highfield, 2014; Penney & Dadas, 2014). Even footage of events is either omitted or self-censored, as activists are aware that they can be reviewed by the police that could lead to charges after the event has finished (Croeser, & Highfield, 2014; Penney & Dadas, 2014).

The activists' increasing concerns over surveillance of social media demonstrated a prescient understanding of the limitations of using Twitter and Facebook (Croeser, & Highfield, 2014; Penney & Dadas, 2014). Activists voiced their concerns about

censorship across numerous social media platforms, especially Facebook. As a result, some posts made on Facebook have been removed by a high proportion of nearly 60% according to a participant, even if they are constantly re-posted (Penney & Dadas, 2014). It is not just Facebook that censors posts - Twitter does too, many activists according to Penney & Dadas (2014) censorship was suspected at various Occupy events (Croeser, & Highfield, 2014). Even though activists are allowed a space on Twitter and are not subject to evictions and daily harassment which makes it difficult to occupy the physical space, use of Twitter still is constraint, as ones knows that space is policed (Penney & Dadas, 2014). Some users have seemingly moved from an open space to a closed space that is known as being the 'Dark Social' (Madrigal, 2012). These spaces are private and potentially unmonitored where they can exchange messages without censoring and (they hope) monitoring (Madrigal, 2012). Even if the Government were to chooe to shut down the social media networks it may have no effect on disorder, as media coverage through the TV and newspapers is still prevalent.

In addition to physical and digital activism, it appears that a new form, labelled analytic activism will become evident in time (Karpf, 2012). This constitutes an approach focusing on citizen-driven politics whereby a new generation of organisations use the Internet to listen their supporters in new ways (Karpf, 2012). Milan & van der Velden (2018) uses a similar term, data activism, to describe this. Analytic Activism seems to focus on the importance of using digital tools to listen, refine strategies and tactics to support the decision-making process (Karpf, 2014 & 2018). It embraces the culture of testing that can guide organisational learning, workflow and practices. This primarily focuses on the analytics rather than the digital conversation. It allows the user to gain insight into public opinion regarding the event (Karpf, 2014 & 2018). Traditional techniques for activism are not entirely replaced, new things do come along, but it can be complimentary. Analytic Activism is something that needs to be taken into consideration by the LEA if it is not already doing so, as it may be seemingly more difficult to police public order events because of this new way of organising the protester organisation group (Karpf, 2014 & 2018). Police are readily analysing events, but activist organisations are not yet actively doing this it seems.

The use of activism analytics may help to be more tactful channel their voice to their perspective audiences, providing greater credibility and enhance their reputation, perhaps help gain greater support for activists (Karpf, 2014 & 2018; (Milan & van der Velden, 2018). This could help channel their protests or demonstrations in a good or bad way. In a bad scenario, an analytic group could extract and analyse the protesters/demonstrator's posts to reshape the story at a faster pace, thereby inflaming the situation and disrupting the police's operations (Karpf, 2014 & 2018; (Milan & van der Velden, 2018). Enhanced intelligence may improve strategic moves that can be made at the event, but this remains nascent, and the police are well-equipped in their analytic capability (Karpf, 2014 & 2018;

Milan & van der Velden, 2018). Though there are negatives to analytic activism, it can also be a tool used for the public good to help maintain the peace by reaching out to people in/outside the demonstration or protest reduce the likelihood of disorder (Karpf, 2014 & 2018; Milan & Gutierrez, 2015). This may keep the focus of the narrative on the topic of the event, raising the awareness for that cause rather it being about something else entirely that is seemingly less relevant to the event (Karpf, 2014 & 2018; Milan & van der Velden, 2018). Furthermore, having this analytic capacity can help spread the message in a democratic and civil way to enhance their campaign. This form of analytic capability poses different challenges and ethical problems (Milan & van der Velden, 2018; Karpf, 2012; Syracuse University, 2014)

2.4 Text and Data Mining

Text mining analyses large amounts of natural language text and detects lexical or linguistic patterns to extract actionable insights from the text to help answer specific research questions (Berry & Linoff, 2011; Han, 2011; Jo, 2019). The data identified is retrieved through an enhanced information retrieval where keywords are searched for to retrieve relevant electronic documents. Text is a common way to exchange and communicate factual information or opinions (Berry & Linoff, 2011; Han, 2011; Jo, 2019).

Social media datasets are large and without automated processing for analysing this data, social media data analytics becomes an unfeasible prospect (at least within a reasonable timeframe) (Berry & Linoff, 2011; Liu, 2015; Han, 2011; Jo, 2019). Text mining has the capacity to automatically help filter large amounts of social media data, extract and organise the relevant text through learning how to find information in each document and analyse its content that contains unique language abbreviations, words that could be either a noun, verb or preposition, codes and symbols (Berry & Linoff, 2011; Liu, 2015; Han, 2011; Jo, 2019). For example, lexical analysis is performed with the aid of domain-specific dictionaries in a form that can allow a computer to extract useful structured information from the original unstructured data source of documents. There are two main approaches to text mining, which are semantic parsing and bag of words (Berry & Linoff, 2011; Liu, 2015; Han, 2011; Jo, 2019). Semantic parsing is based on word syntax (with focus on word type and order) and can create many features on a single word by tagging it, such as a noun, named entity or be tagged part of a sentence, so this single word could have three features (Berry & Linoff, 2011; Liu, 2015; Han, 2011; Jo, 2019). The bag of words method does not consider the word type or order, rather it takes words to be attributes of the document and treats each word as a single token in a sentence.

In addition to researching features of the text, text mining has additional functions, as it can identify whether a person = human, happy = emotion, and television = object.

The text mining workflow from problem definition and goals, through identified text collection, text organisation and feature extraction (extracting word tokens into matrices or calculating sentiment) to analysis, where patterns and trends are identified across millions of documents with the aid of semantic verification and data visualisations to find new knowledge (Berry & Linoff, 2011; Liu, 2015; Han, 2011; Jo, 2019). For example, it could identify that most people were happy at 2pm on a specific day of a demonstration. This new knowledge may provide a series of insights or could help to provide a set of recommendations (Berry & Linoff, 2011; Liu, 2015; Han, 2011; Jo, 2019).

Text mining requires the use of many different tools and techniques, such as natural language processing, information extraction, information retrieval, text categorisation, text clustering, document similarity, document frequency and summarisation (generally sentence-based or keyword-based), sentiment analysis and data mining to help turn the data into knowledge (Berry & Linoff, 2011; Liu, 2015; Han, 2011; Jo, 2019; Witten, 2005). Text mining extracts clear information explicitly from the text whereas data mining implicitly extracts information that is previously unknown and/or hidden in the input data. The use of both techniques could lead to gaining detailed insights into answering a research question.

Data mining is the exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules. In data mining, hidden, unnoticed patterns and trends are investigated (Berry & Linoff, 2011). Data mining is known as the analysis stage of the Knowledge Discovery and Data Mining (KDD). KDD is a field of computer science that includes the tools and theories to help humans in extracting useful and previously unknown information (e.g. knowledge) from large collections of digitized data. KDD is the process of discovering useful knowledge from a collection of data (Berry & Linoff, 2011). Data Mining is the application of a specific algorithm or algorithms to extract patterns from data. KDD processes are outlined in Figure 2.3:

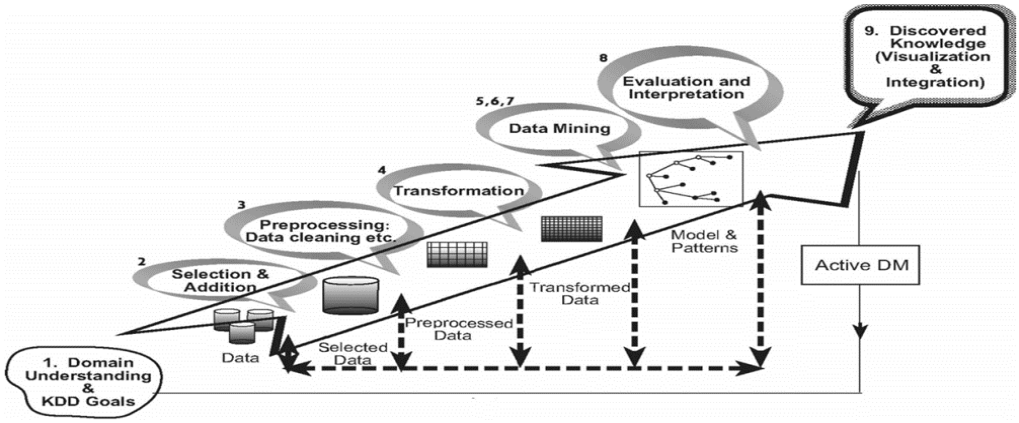

Figure 2.3 KDD Process (Maimon & Rokach, 2005)

This process has several steps (Berry & Linoff, 2011; Liu, 2015; Han, 2011; Jo, 2019), which are: -

- It starts with developing an understanding of the application domain and the goal and then creating a target dataset.
- The text mining phase that cleans and reduces data and outputs the transformed data.
- The next step is using data mining to identify patterns.
- Finally, discovered knowledge is consolidated by visualisation and interpreting any results or patterns found.

The steps display in Figure 3 demonstrates that KDD is extracting knowledge from data while data mining is an application of specific algorithms that identifies patterns (Berry & Linoff, 2011; Liu, 2015; Han, 2011; Jo, 2019). Data mining is an interdisciplinary field where and brings together researchers from several different fields such as statistics, mathematics, artificial intelligence, machine learning (which is a branch of artificial intelligence), sentiment analysis (branched under data mining) clustering algorithms, data visualisations and databases (Berry & Linoff, 2011; Liu, 2015; Han, 2011; Jo, 2019).

2.5 Social media data mining

Social media data mining as an area is vital to understand its current/ future development of social mining techniques to determine which techniques are the most effective working with social media and social network data (Berry & Linoff, 2011; Liu, 2015; Han, 2011; Jo, 2019). It is important to note that technology and its social applications have provided a new way of communicating, and the public have begun turning to this as a first sensor. This enables researchers to observe human behaviour differently at an unprecedented scale (Castells, 2009; Chui, 2012; Greer, 2010). These vast quantities of social media data are being produced every day in real-time, and they are unmediated, rich and interlinked. Social media data is in between the real world and the virtual world and requires the use of a combination of sociological and computational methods. This allows researchers to study how individuals share, interact and form communities (House of Commons, 2014b). This has given researchers and other organisations an opportunity to mine human behavioural patterns, learning more about the ways in which to understand about how people and society work. For example, social network analysis could help to identify those at high risk for involvement in violence or aid to identify individuals at high risk for being involved in violence (Hollywood et al., 2018). This additional knowledge may enable improvements in the design of computing systems to improve public safety online.

Social media data is unstructured in different text data formats (HTML, XML, JSON & CSV) and of a low quality. The collected data is not representative of the population (Batrinca & Treleaven, 2014; Boyd & Crawford, 2012; Fan & Bifet, 2013; Vis, 2014). There are many data quality issues, ranging from missing values, inconsistent values, verification of data processing and completeness of near real-time data, relevancy of the data and limitations of tools for analysis. Text classification is important in many

application domains, as there is a higher volume of short text from social networks and online review systems. Text mining traditionally has limitations when automatically classifying short text, as a result of sparseness, informal sentence expression and lack of context. Text classification techniques usually take advantage of information redundancy in a well written document, which occurs more often with long documents than short ones (Dai, Sun & Liu, 2013). This complex nature of social media data makes it challenging to make meaningful observations from the data sourced. We will see further examples of this in section 2.5.

The application of data mining methods to social media is relatively new compared with other areas of study in social network analysis despite from research dating back to the 1930s (Aggarwal, 2011). These challenges to the data mining field led to discovering that existing methods that are effective in the analysis of social media data together with the use of sociological theory to adapt these techniques to investigate the data allow researchers to gain a higher level of understanding. Preotiuc-Pietro (2014) and Weller et al. (2014) have outlined several existing data mining techniques e.g., K-means (finding the centre of a cluster) that apply well with social media data. For instance, clustering techniques have been used to find the most central point of diffusion of the protests to the central point of the tweet located via geo-location of the tweet and hashtag (Bastos et al., 2014). This helped to understand that users tweeting about street protests and users in geographically isolated areas relied on Twitter hashtags to engage in the demonstrations (Bastos et al., 2014).

2.5.1 Police and Social Media Data Mining

Currently, some police forces are using SAS Enterprise Miner and SPSS Modeller data mining tools are being used to improve intelligence. In the US, Richmond police use SPSS and other tools, to analyse changing drug patterns and trends (McCue, 2003, 2006 & 2010). SAS is used by some police forces in the UK, notably West Midlands Police (WMP), both to correct and match aliases or false birth dates in the database records to improve the quality of their information, as previous used systems lacked these characteristics (SAS, 2015a).

HMIC recognises that advanced technical methods now exist to develop a data-mining engine to scan social media for signalling crime, disorder and control to identify potential disorder and crime (HMIC, 2011a & 2011b; HMICFRS, 2021). These methods make it possible to detect patterns and trends in behaviours and events that deviate from normal events that have been previously monitored (HMIC, 2011a; HMICFRS, 2021). Key training must be provided to relevant staff to make sense of the data gathering (Wyllie, 2013; HMICFRS, 2021) to support police services to prevent crime and ensure public safety. Data mining algorithms' can discriminate, which needs to be accounted for when making decisions on police strategy (Essers, 2013; HMICFRS, 2021; House of Commons, 2013 & 2018; Kumar, 2013).

2.5.2 Sentiment Analysis of Social Media

2.5.2.1 Background on sentiment analysis and police context

Sentiment analysis detects emotion in text, such as opinions on a topic, where it mines the emotions, feelings and attitude. Droba (1931) suggested this area of studies was measured and quantified from questionnaires in 1931, and Public Opinion Quarterly (2018) stated that a scientific journal on public opinion was established in 1937. Since 2004, research into this field has been steady, and there has been a rise in the popularity of sentiment analysis for detecting emotions on opinions (Mäntylä, Graziotin, Kuutila, 2018). This is due to the rise of product reviews on the web in the after 2004 and as of 2018 there have been close to 7,000 published papers in the area undertaking work such as analysing social media posts to predict financial markets rise or fall (Mäntylä, Graziotin, Kuutila, 2018). Sentiment analysis can be divided into different subtasks that can analyse text, such as (Westerski 2008): -

- *"Sentiment context—to extract opinion, one needs to know the 'context' of the text, which can vary significantly from specialist review portals/feeds to general forums where opinions can cover a spectrum of topics."*
- *"Sentiment level—text analytics can be conducted at the document, sentence or attribute level."*
- *"Sentiment subjectivity—deciding whether a given text expresses an opinion or is factual (i.e., without expressing a positive/negative opinion)."*
- *"Sentiment orientation/polarity—deciding whether an opinion in a text is positive, neutral or negative."*
- *"Sentiment strength—deciding the 'strength' of an opinion in a text: weak, mild or strong."*

These different techniques can be applied in different situations, for instance, Liu (2012, 2014 & 2020) said that sentiment analysis can be used to identify a brand's influencers and promotors, and that may allow the application of sentiment orientation to source negative content against the organisations brand. This enables the organisation to engage with the individual(s) to improve their perception. Sentiment analysis accuracy can in some circumstances be questionable, but new research to improve techniques is continuing to enhance the accuracy rates, thus improving the output of social media monitoring (Liu, 2012, 2014 & 2020). On-line opinions are valuable for businesses when identifying new opportunities, improving on marketing products and cultivating their digital reputations (Liu, 2012, 2014 & 2020). As outlined in section 2.3.2.3, it can be difficult for companies to identify the most valuable consumer posts.

A Sentiment Analysis (SA) tool may be useful for the police to monitor the level of tension expressed from the moods and emotions on social media (Liu, 2012 & 2015). For example, the police might use SA to detect impending disorder using key terms that may be associated with social, political, economic, and environmental subject areas within the community (HMIC, 2011a; HMICFRS, 2021; House of Commons, 2018). Early detection of rising tension may help the police to act quickly in opening-up effective two-way communication channels to the most disaffected, alienated and vulnerable (HMIC, 2011a; HMICFRS, 2021; House of Commons, 2018; Institute of Community Cohesion, 2010). This may enable the police to give firmer reassurance to the social media community that they are working hard to eradicate the underlying cause of tension to maintain the peace (HMIC, 2011a; Institute of Community Cohesion, 2010). A good tension indicator "offline" is one of the main priorities for the Association of Chief of Police Officers (ACPO), as it seems to be a key aspect to their policing decisions (King and Waddington, 2004):

"A good tension indicator or community monitoring system is seen as the key to providing the police with advance notice that trouble may be brewing and giving them adequate opportunity to work in conjunction with community partners to minimize the risk of disorder (ACPO, 2000: 10)." (King and Waddington, 2004, p.121)

There is a need "offline" for a tension disorder model, but with modern technology and social media there seems to be a necessity for similar methods to be applied "online" to help the police minimise disorder (HMIC, 2011a; Institute of Community Cohesion, 2010). The SA approach can help to collect data unobtrusively and be scalable to measure an understanding of the emotional impact of an event (Liu, 2012 & 2020; Medhat, 2014). During the past 10 years, the police have in some capacity been using sentiment analysis, but only in about 10% of their work duties (Dencik et al., 2015 & 2017). Algorithmic intelligence on sentiment for policing purposes is therefore still not a major practice yet, although it may become so as the level of sophistication of algorithms increases to serve a purpose when informing real-time police tactics (Dencik et al., 2017; HMICFRS, 2021). In addition, Dencik et al. (2015, pg. 28) outlined that it is important with this type of analysis that one have *"require contextual knowledge of language that can account for different demographics, places and cultures."* The use of technology with local human contextual knowledge is important to assess the analysed data when making a professional judgement on adapting tactical position for an event (Dencik et al., 2015). This human control can highlight the role of discretion in the use of predicative analytics; without this, mistakes may be made since there is a margin of error within data analysis (Dencik et al., 2015).

2.5.2.2 Sentiment analysis approaches

Sentiment classification techniques can be divided into machine learning approaches, lexicon-based methods and a hybrid style. The machine learning method uses linguistic features and algorithms to classify the data (Medhat, Hassan & Korashy, 2014). The lexicon-based approach is dependent on a sentiment lexicon, which contains a list of weighted sentiment terms as scores, such as +1 or -1 (Liu, 2020). This is subdivided into dictionary-based or corpora-based methods that applies semantic and statistical methods respectively to identify sentiment polarity (that is, whether the sentiment is either positive or negative) (Liu, 2020). The hybrid methodology combines both the machine learning and lexicon-based methods.

The machine learning based approach uses classification techniques such as Maximum Entropy and Neural Networks to classify text (these algorithms will be further explored in section 5.9), splitting data into a training and a test dataset (Liu, 2020). The training set is used to learn the differentiating characteristics of a document, while the test set is applied to check the performance of the classifier (Liu, 2020). The benefit of this method is that one can create and adapt trained models for different contexts and purposes. The features of machine learning based approach for sentiment classification are (Liu, 2012, 2014 & 2020): -

- Term existence and their frequency, which includes n-grams e.g. unigrams and bigrams;
- Part of speech information is used for disambiguating sense that guides the feature selection;
- Negations can be difficult to detect, but there is a chance of reversing a sentiment opinion that expresses positive or negative sentiment.

These key features are represented as feature vectors that are used for the classification algorithm. The text classification methods that use the ML approach can be divided into supervised and unsupervised learning methods (Liu, 2012, 2014 & 2020). The supervised methods make use of many labelled training documents. The unsupervised methods are used when it is difficult to find these labelled training documents (Liu, 2012, 2014 & 2020). According to Appel, Chiclana & Carter (2015) the main difference between both supervised and unsupervised approaches is that 'supervised learning' uses classification techniques relying *"on the training set used, [and] the available literature reports detail classifiers with high accuracy"* (Appel, Chiclana, Carter, 2015, p.g.7) However, these tend to be tested on a limited kind of sentiment source, such as film reviews (Appel, Chiclana & Carter, 2015). This limits the performance of indication of sentiment in other general cases. Furthermore, if training set is limited this may cause algorithmic bias which could affects the machine learning process and lead to a potential negative outcome, which is an area the UK police and other industries need to improve upon (Babuta & Oswald, 2018; CDEI, 2020a & 2020b;

HMICFRS, 2021). Unsupervised learning uses sentiment driven pattern to acquire labels for phrases and words. However, ML is limited as emphasised with issues related to algorithmic bias, so it is not a substitute for the human brain, because there is less flexibility when inferring outside the parameters what has been learnt (Cambria et al., 2017; Liu, 2012, 2014 & 2020). Therefore, it is important to test the algorithm to ensure that substantial issues are not caused in a real-world context (Babuta & Oswald, 2018; CDEI, 2020a & 2020b; HMICFRS, 2021). The main advantage of ML approach is its ability to adapt and create trained models for *specific* contexts and purposes. The limitation is the difficulty of integrating this into a classifier, as the level of generalisation might not be acquired from the training data used (Cambria et al., 2017; Liu, 2012 & 2014).

The lexicon-based approach is dependent on locating (or constructing) an opinion lexicon that is used to analyse the text (Cambria et al., 2017; Liu, 2012, 2014 & 2020). In each of the dictionaries there is a set of words, where each word has been assigned a calculated polarity score. There are multiple methods to construct a lexicon, for instance, one possibility is to find seed words that define two poles of semantic axis, such as good or bad and to search a dictionary for synonyms and antonyms of these (Cambria et al., 2017; Liu, 2014 & 2020). A second possibility is a corpus-based approach starting with a seed list of words and finding opinionated words in a large corpus to locate words with orientations that are context specific. This could be achieved using semantic or statistical methods (Cambria et al., 2017; Liu, 2014 & 2020). Thirdly, manual construction can be applied, but this is difficult and time-consuming, as it requires specialist linguistics. There are many dictionaries (e.g. SentiWordNet, SenticNet, Stanford and SentiStrength) that are based on the English language, but most are American English than UK English (Cambria et al., 2017; Liu, 2014 & 2020). Furthermore, other dictionaries with different languages are sparse in comparison with English based dictionaries (Cambria et al., 2017; Liu, 2014 & 2020). These dictionaries may have a wider term coverage, but there are a comparatively limited number of words with a fixed sentiment orientation or score assigned to the words (Cambria et al., 2017; Liu, 2014 & 2020). Additionally, set words included in the dictionary could be context-specific, which may not have the same meaning for another topical event.

Lastly, the hybrid approach combines both the lexicon-based and machine learning approaches, which has the possibility to enhance the sentiment classification performance (Cambria et al., 2017; Liu, 2012, 2014 & 2020). One or more of the sentiment dictionaries can be used for initial sentiment detection, and then labelled items can feed directly in a series of ML techniques, such as Naïve Bayes and Support Vector Machines (Cambria et al., 2017; Liu, 2012, 2014 & 2020). The benefit of a hybrid approach are the lexicon's symbiosis, detection and measurement of sentiment at the concept level and less sensitive to changes in a topic domain (Cambria et al., 2017; Liu, 2012, 2014 & 2020). These sentiment analysis approaches will be discussed in greater

depth in the social media strategy section 3.1.6, where one will be selected to be implemented in the project

There are series of lexicons for sentiment analysis that use different techniques and some of these are outlined in Table 1.

Sentiment analysis dictionaries for analysis	**Employed techniques by the dictionaries**
Affective Norms for English Words (ANEW)	A lexicon that provides a series of normative emotional ratings for 1,034 English words (Bradley & Lang, 1999)
AFINN	A revised ANEW version focused on language used in microblogging platforms which forms AFINN (Nielsen, 2011).
LIWC	The approach analyses positive and negative sentiment, but it also includes cognitive, emotional and structural components of text. This uses a dictionary that contains words and their classified categories (Pennebaker, Booth, & Francis, 2007).
SentiStrengh	This sentiment lexicon uses assigned scores for positive and negative phrases in text (Islam & Zibran, 2017; Thelwall, 2019).
SentiWordNet	The SentiWordNet is an extension of the English lexical dictionary called WordNet that gathers nouns, verbs and adjectives into synonym sets called synsets (Baccianella, Esuli, Sebastiani, 2010; Esuli, & Sebastiani, 2006).
SenticNet	This natural language processing approach infers polarity at the semantic level (Cambria, Poria, Bajpai & Schuller, 2016)

Table 1 Description of sentiment analysis tools

The different lexicons for analysis outlined may be applied to different subject areas, such as politics, business and public (Andrea et al., 2015; Cambria et al., 2017; Geethaa, Singha, Sinhab, 2016; Kucharska, 2018; Sailunaza & Alhajjab, 2019). The research of the lexicons will be further explored in the initial data and information processing in section 5.9. There have been numerous studies on the area of reviews of products and services that have been critiqued by their customers. There are a

number of other websites that automatically summarise product information and collate these customer reviews. For instance, this can relate to opinions about travel, restaurant reviews and store guide for customers searching within Google and Bing that compute their star ratings (Andrea et al., 2015; Cambria et al., 2017; Geethaa, Singha, Sinhab, 2016; Kucharska, 2018). In the context of sentiment analysis, businesses monitor their brand reputation, competitive research and online advertising (Andrea et al., 2015; Cambria et al., 2017; Kucharska, 2018). There are organisations that monitor social media platforms, such as Twitter and Facebook for their brand, while some may have make use of off-the-shelf products, such as SentiOne (https://sentione.com/) or Clarabridge, rather than developing an in-house solution (Andrea et al., 2015; Cambria et al., 2017; Geethaa, Singha, Sinhab, 2016; Kucharska, 2018). These types of tools will be further explored in the social media research strategy in section 3.1.4. Online advertising is a major source of revenue and sentiment analysis applications have been used within "Blogger Centric Contextual Advertising", which highlighted dissatisfaction with personalised adverts in a blog page (Andrea et al., 2015; Cambria et al., 2017). In terms of politics, Governments appear to reach out to the electorate to receive voting advice on policy, and gauge sentiment based on public opinion (Neuropolitics, 2016). As a result, this can help to contribute towards an understanding of how the electorate feel about different issues relating to speeches and actions of each political candidate or Member of Parliament (MP) (Neuropolitics, 2016). In these examples, there are different challenges with their approaches, especially with respect to social media. For example, the ever-evolving nature of (the English) language and having to express a view within a short space presents difficulties (Liu, 2012 & 2014). Some of these challenges have previously been outlined, but there are other important considerations to take into account (Liu, 2012 & 2014) such as: -

- Spelling mistakes or texting language where words are shortened intentionally can make it difficult for the classifier to detect and classify the words. The words that are not spelt in their normal convention, will require replacing with the correct spelling or be added to the dictionary.
- Calculation of emotional valance of each sentence is an issue that is widely recognised in the Sentiment Analysis field.
- The assigning of values to words can be difficult and at times be inaccurate, especially when dealing with sarcasm, slang, irony and idioms.
- Different lexicons' dictionaries label the same words differently in terms of their sentiment whether it may be positive, negative, happy or sad depending on the context in which they have been aligned.
- Some algorithms may be suited to the classification of short text over longer documents, while some may require more words to give a higher rate of accuracy.

- Splitting sentences can be done incorrectly depending on how an algorithm interprets the sentences.
- Negation is not reserved and uses a negating function to calculate the sentiment value based on collection of phrases containing negating verbs and adjectives (Liu, 2020). Intensifiers refer to words, such as 'quite', 'most' and 'extremely' that change the sentiment when adjacent to non-neutral terms, such as 'wrong' and 'happy'. These intensifiers are divided into two categories, namely amplifiers (most, extremely) and down-toners that increase or decrease the intensity of the sentiment by a set percentage (Liu, 2020).

Even though there are many technical challenges to overcome, researchers, businesses and organisations continue to strive for new techniques (or to combine together existing methods) to achieve higher levels of accuracy and representativeness in sentiment analysis (Liu, 2012 & 2014).

2.5.2.3 Previous sentiment analysis studies on Twitter

Twitter has been used for sentiment analysis in many studies, of which most are in non-security domains, such customer reviews of hotels, user reviews on products and feedback based on box office movies. In particular, the tourism domain (Flores-Ruiz, Elizonso-Salto, Barroso-Gonzalez, 2021; Garcia, Gaines, Linaza, 2012) has introduced the use of lexicon databases for sentiment analysis of user reviews sourced from TripAdvisor regarding food and accommodation. In addition, social media data is used to support studies (Xu, Zhu, Bellmore, 2012) into bullying by using text classification to identify various emotions, such as empathy, sadness, pride and anger in tweets. In another project (Mittal, Goel, 2013), Twitter was used to understand the difference between market and public sentiment, where text classification was applied to classify sentiment into four different classes: happy, kind, alert and calm. This was used to identify previous Dow Jones Industrial Average changes in order to subsequently predict future stock fluctuations. These examples show that social media and the application of sentiment analysis to social media data has been applied in different contexts. There have been a series of advancements with the combination of intelligent systems and social media analysis designed for decision-making relating to public safety. Dencik, Hintz & Carey (2018) and Glass & Colbaugh (2011) have addressed some issues related to security with predictive analysis and situational awareness. Glass & Colbaugh (2011) proposed a methodology to evaluate real-world events with the use of a Violence Detection Model (Cano et al., 2013) to locate violent topics discussed on micro-blogs. Furthermore, during the Great Eastern Japan Earthquake (Sakaki, Toriumi, Matsuo, 2011), social media posts were analysed to identify the relation between people's activities and what happened after the event transpired.

The research conducted on analytical capability of big data and social media data of security and (dis-)order events is limited and requires greater research, such as the dynamics of institutional application, interactions between data analysis and human intervention (Dencik, Hintz & Carey, 2018). There have been different lexicon-based sentiment analysis algorithms used in various situations, such as detecting radicalisation in social media (Bermingham, 2009; Cohen et al., 2014), where social network and lexical analysis were used to identify and understand the characteristics of radicalised users. This analysis showed that there could be a way to identify lone wolf terrorism. In our research project, there is a focus on the potential disruption of public order at demonstrations. There has previously been some research using Twitter data based on demonstrations, such as Jurek, Mulvenna, Bi (2015) who used sentiment analysis to improve lexicon-based-sentiment based on a series of English Defence League (EDL) UK demonstrations. This analyses the sentiment of Twitter posts related to the EDL and level of (dis-)order during the event. A lexicon-based approach is adopted but the researchers noted a drawback of using an English dictionary as users participate around the world (Jurek, Mulvenna & Bi, 2015). Therefore, these authors decided to translate the language of the sentiment lexicon while making an application of sting similarity functions. The authors used SentiWordNet as a baseline to then manually create a sentiment lexicon of 6300 words for the context of demonstrations.

The focus of Jurek, Mulvenna, Bi (2015) was on the relationship between public sentiment and the tension of the EDL event, and whether it could be used to predict the level of disruption. The lexicon applied was reduced from 6000 to 1500 words, as the focus was negative sentiment based on the violence and disorder through the event. The most negative of five EDL events was located in Birmingham had the highest level of disorder and arrests. The tweets prior to this specific event had a level of negativity three times higher when compared with a similar event in Brighton, which had a peaceful event. The research suggested the results are useful as an indicator for the level of disorder, which could be used by the police for planning resources to safeguard events and the use of sentiment analysis for prediction and monitoring of events.

Bahrami et al. (2018) agree with Jurek, Mulvenna, Bi (2015) as to the use of predictive capability, but this paper is set in the context of an American protest, where event-specific features were used for prediction purposes, which notably heightens the level of accuracy but is perhaps unrealistic to use in more general situations. Williams et al. (2013) use sentiment analysis to monitor the level of tension on social media to identify deviations from norms in terms of low-level tension. The information is used in combination with statistics based on deprivation, demography and neighbourhood crime to provide a more complete view of both physical and online events. The resulting outcomes provide neighbourhood informatics of the event to answer questions regarding civil unrest to help keep the public safe. This research was

provoked by the HMIC (2011b) report as it outlines a way for using technology for anticipating future public disorder when assisting police to further their understanding in an offline-online operational setting. This tension monitoring application requires further work. As outlined, it has been tested on only one dataset and aims to consider the *"reciprocity between online expression and offline action".* As Williams et al. (2013), Bahrami et al. (2018) and Jurek, Mulvenna, Bi (2015) outline there is a way to potentially use a predictive capability to assist the police to keep the public safe. As Dencik et al. (2015) suggest, algorithmic intelligence is not yet a major practice within LEAs, but it's uptake will expand once the level of sophistication, transparency and reliability of an algorithm increases (CDEI, 2020a & 2020b; HMICFRS, 2021; Kearns & Muir, 2019). These papers highlight that if greater progress is made in sentiment analysis and the overall general predictive capability, then there will likely be a wider uptake of this being applied within the LEA and sentiment analysis of social media data and its application in the context of UK policing is a developing area that requires greater research.

2.6 Summary

The aim of this literature survey was to review different studies on police practice, public order, application of social media in LEAs and social media audiences. Additionally, the literature review focused on the evaluation of text mining and data mining with a focus on sentiment analysis of social media in a security context. These topics are important in understanding the knowledge domain, synthesising knowledge and rationalising its significance.

The research showed the LEA historical context and current developments on the use of technology, public order approaches and the application of social media data analysis in the context of sentiment analysis to assist in the police operations for public safety. There are a range of examples of how sentiment analysis has utilised Twitter data for reviews on products and services, public safety when earthquakes occur and security of demonstration and protest events as a predictive capability (early warning detection for tensions) to assist the LEA in maintaining the peace.

Social media in general, and Twitter in particular, is one of the main communication tools for organising and providing information on demonstrations and protests and analysis of the available data can be used to inform the public of existing/ new developments. Therefore, it is important to research this platform's content to understand the public opinion on a demonstration and to understand the demonstrations themselves. The literature review has identified that various studies on public order have been qualitative rather than quantitative insights, which could be due to a shortage of technical skills in the academic community. While there are projects that conduct quantitative research, such as Williams et al. (2013), these are comparatively few. Moreover, the research shows a greater need for further research

on sentiment analysis of social media in the context of UK demonstrations and the application of social media data mining.

The literature review shows both police use of social media/ social media data in public order events, social media data mining, and the application of sentiment analysis are continuously changing landscape in the field. Research suggests that public order events applied in the use of sentiment analysis of social media data in the UK requires further investigation due to limited research in this area. The literature identified a series of projects (Bahrami et al., 2018; Flores-Ruiz, Elizonso-Salto, Barroso-Gonzalez, 2021; Garcia, Gaines, Linaza, 2012; Jurek, Mulvenna, Bi, 2015; Xu, Zhu, Bellmore, 2012) that have applied sentiment analysis of social media are more often using one or few dictionaries to detect the tweets sentiment rather than the application of machine learning or a hybrid approach (refer to both the list of algorithms in section 5.9 and sentiment analysis approach in section 3.1.6). This project will focus on a wider range of dictionaries, and algorithms to identify which is strongest on public order, along will be further discussed in data mining approach in section 5.9.

Both the literature review and feedback for the publication (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018) will help inform the social media research strategy in the next chapter that will define the framework to be implemented. This will help to address the socio-technical aspects of the problem, as social media data is qualitative data on a quantitative scale (D'Orazio, 2013).

3 Social Media Research Strategy

This research has drawn from social science methods and applied it to a series of specific of events in the form of a case study, which has been generalised to relate each of other public order events (refer to section 3.1.3 for selected datasets) with the use of computational methods (refer to sections 3.1.4 to 3.1.6) to explore data on a wider scale. In the traditional approach a smaller sample of data will be of focus, but with computation techniques this sample can be on a larger scale. Furthermore, the largest demographic (refer to literature review in section 2.3.3) that appear to both use social media and attend these events, seem more relevant source of social media data on platforms such as Twitter.

Section 3 discusses a series of social media research strategies and how social science and computational methods are integrated into our proposed social media lifecycle. This first section and section 3.1 of the chapter have been published in the conference paper (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018), which is provided in appendix 10.9 and from which we draw from throughout. A pilot study was undertaken to inform future practice and to adapt the social media lifecycle from a business to research context.

Upon reviewing a wide range of papers, it was noted that some provided an excellent, thorough description of the steps they took in their research. However, it was often found that the initial stages of the research that would be needed for a complete addressing of any research question were poorly defined. The available literature tends to be project-specific in its approach and is therefore not immediately suitable for generalisation to other research, which is not unexpected, given that social media research methodology is a topic still in its infancy. It can be difficult for any given researcher to know where to start in the area and to identify what decisions need to be taken to form a social media methodology for the project in question.

The research community and other organisations are trying to come up with better ways to express their social media strategies, such as the SMDS project, which *"focuses on studying practices behind and attitudes towards the collection, storage, use, reuse, analysis, publishing and preservation of social media data"*. SMDS has produced a social media data process that aims to clarify for researchers the layout and order of each phase that may be required in a social media data project. This focuses on the data management process of social media data and aims to help researchers to consider their attitudes towards the data they wish to work with. What we aim to do in this chapter is to identify a complete set of stages for any social media research project lifecycle to follow, including within this the SMDS insights into data management, as these touch on highly pertinent points within the overall process.

Having found the nascent SMDS data management paradigm, we continued the search for a full social media project lifecycle. While this proved impossible to source as no such lifecycle yet exists, we did encounter a somewhat developed social media research project lifecycle created by the UK Government Social Research (GSR) service. The GSR based its lifecycle on the Cabinet Office framework for data science projects, as it had *"numerous parallels here"*. This lifecycle has been tested on two social media projects within Government, namely: using Twitter to predict cases of Norovirus and assessing the experiences of the 20th Commonwealth games held in Glasgow, with reports on the analysis of broadcast and online coverage being produced using the strategy. There is no publically available information on whether or not this social media lifecycle was in fact a success. However, GSR produced outcomes that may be a measure for potential successes. For example, the Commonwealth games on Twitter were in the top 10 highest sporting event hashtags of the year, generating a highly positive contribution to Scotland and Glasgow both internationally and within the rest of the UK. Furthermore, GSR identified that between 14/06/14 to 06/08/14, there were 3.2 million mentions of the Commonwealth Games on social media in the English language. There were other positive outcomes, which enables GSR to identify where future improvements can be made with the organisers in raising the profile for relevant cities and events. In the sequel, we shall aim to integrate aspects of the GSR service lifecycle and the SMDS data management process alongside our own insights into the social media project lifecycle.

3.1 Our integrated social media project lifecycle

The GSR social media project lifecycle consists of seven stages: Stage 1: Rationale – Business/Citizen Need, Stage 2: Data, Stage 3: Tools and Output, Stage 4: Research Phase, Stage 5: Implementation/Publication/Action, Stage 6: Evaluation and finally Stage 7: Business as Usual. While this is a useful basic framework that will help to guide researchers through their social media projects, it still requires further development and refinement as the considerations outlined at each stage are given in little detail. Furthermore, this lifecycle was designed to be applied in a commercial and governmental context, which can make it difficult to know what to do at each step from a research perspective. Nevertheless, we have chosen to adopt this framework as a starting point as it proved itself helpful in structuring our own initial social media research project. The research we are conducting aims to enhance the analysis of social media in the context of public (dis-)order events. We investigate how social media data are stored (big data issues), collected, analysed (text mining and sentiment analysis) and then disseminated (to the police, to help predict when disorder may occur). This will form part of the creation of a model to analyse social media data to try to predict the escalation of such events. We will adapt the GSR lifecycle to suit the needs, aims and goals of research projects (as opposed to governmental projects), and a diagram showing the relevant adaptations is displayed in Figure 3.1.

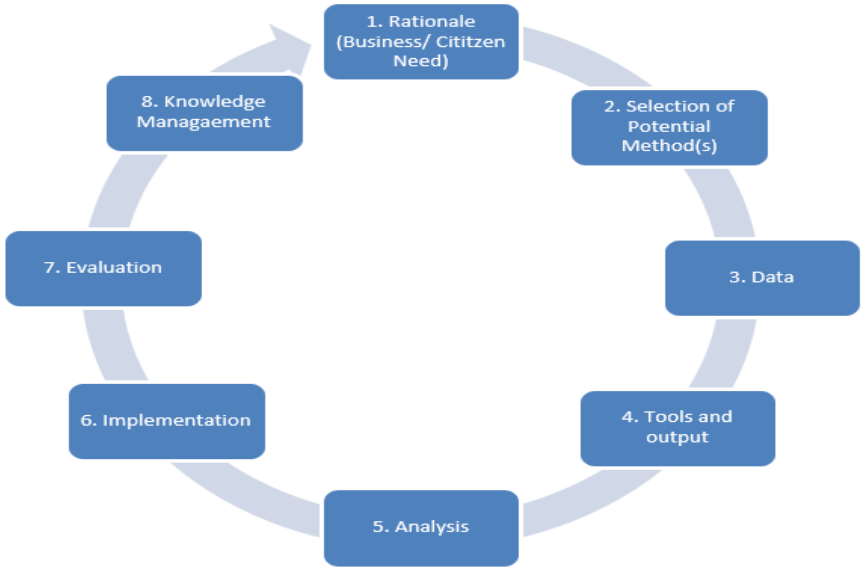

Figure 3.1 Social media research project lifecycle

The steps in the lifecycle are explained below. We will outline the purpose of each step and show where modifications have been made to the GSR lifecycle. The lifecycle explained below will be informed by the pilot study we conducted, which has involved analysing Twitter data around the time of the Baltimore riots (refer to background information on this case study in section 3.1.3.1), with the aim of developing models to identify potential riots before they occur.

1) Stage 1 (Rationale – Business/ Citizen Need) is described as a need to think about social media's attributes (e.g. speed, cost, real-time production). On the basis of these attributes, there are suggestions for the business or citizen's need to be based on: *"using insight to deliver a more timely service to the citizen with fewer resources through the support of social media analysis than would have been possible with traditional means."*. To measure if the project is delivering a timely and resource efficient service to the citizen can be difficult to determine in some cases without actually conducting the project. A rationale for the research must be established, as without this the project will likely lack focus and be too broad, weakening any results or insights obtained. This means that valuable resource that could potentially be better utilised elsewhere is being wasted. While nothing new has been added to this section compared to the GSR lifecycle, we have placed it into the appropriate research context. This stage in our process is important, as one must have a question to drive the collection and analysis of data in research and, one should not let the data drive the researcher. Without a suitable research question, the project would lack purpose. The rationale for the project we carried is outlined above.

2) Stage 2 is a new step which has been introduced called "Selection of Potential Method(s)". This step is required to help adapt this commercial lifecycle into a research context where consideration must be given as to which methods (for example, case study or archival research) will be applied in the research process. This must be decided early on in the process, so that the following stages can take this into account when making relevant decisions in the latter phases of the lifecycle. If this step is not undertaken explicitly in a research context then results may be obtained that are of a particular nature, without account having been taken of the fact that the nature of the methods employed is inextricably linked with one's research outputs. This may cause a loss of momentum in the stages ahead, where special account would have to be made for the method or methods employed. For our particular research, we select a case study-based approach to allow us to work with particular disorder events immediately and then attempt to generalise these to the wider public order context.

3) "Data" is now stage 3 of the lifecycle. It is emphasised that *"The primary purpose of this data is not for research so consideration should be given to representativeness, robustness and ethics."* This statement is confusing, as the same level of rigour would apply in a research context. In this section, the researcher must justify the datasets to be used in the project and examine any necessary ethical considerations regarding the use of the social media data in question in their research. The original purpose of this section remains the same as in the original GSR lifecycle. This phase considers which dataset(s) may be explored to answer the research questions of the project. There is extra emphasis on selecting the correct data as cost may well be an issue here, more so than for a government entity, depending on the size of dataset required for the research, given the finite nature of research grants in particular. This step is

also useful in providing time to think carefully about the selection of datasets. If the data are chosen without due care then this will impact the cleaning, analysis and output of the project, though given the emerging nature of social media technology, it can of course be difficult to fully understand the range of data and metadata that are available before one already has a sample to hand. To that end, collection of a small pre-sample of data can also be a useful initial substage here. The dataset used for the pilot study is based on collecting live data from the 2015 Baltimore riots, USA. This pilot study will help to inform the collection of further datasets, on which the pre-processing and data manipulation scripts developed for the Baltimore data can be re-run.

4) Stage 4, "Tools and Outputs" is named the same as in the original GSR lifecycle. In this phase, the use of specialised social media tools can help to make cleaning and analysis of the collected data easier for researchers. Furthermore, social media data may require manipulation to *"render it useful in a social research setting"*. The outputs from analysis of these data can range from traditional reports showing present findings to predictive models designed to solve real time problems. GSR's process for this step is kept, but in addition to this, the researcher must outline their data collection strategy to show how relevant data in relation to any research questions will be obtained, as well as considering how those data will be stored and whether single or multiple platforms are to be used as this will have an effect on the tools chosen. There are a plethora of tools available for data acquisition, processing and analysis and the tools to be used must be selected with care to ensure that they are both suitably secure and efficacious for the data in question, otherwise time will be invested in tools that are not appropriate for large scale data retrieval (not all return the same metadata, for example), cleaning and/or analysis. The tools selected will depend upon the platform from which data are to be extracted. In our case, since we are dealing with Twitter data we chose NVivo NCapture to extract a live sample of data from the Baltimore riots and used R for data manipulation. For the retrospective datasets that we collect in the future, we will instead be using DiscoverText for acquisition. This tool is widely used in the research community because it provides access to one of the cheapest ways to retrieve a complete historical record from Twitter's official provider GNIP. Even though the extraction and analytical tools are being selected at this stage, the actual techniques for analysis will be investigated in stage 5.

5) Stage 5 was originally named "Research Phase" in the GSR lifecycle, rather than "Analysis". Clearly, given that we are aiming to develop a full research lifecycle, the former name is no longer appropriate. This step emphasises that care must be taken regarding the representativeness of data to mitigate any bias in the analysis. Lastly, *"Care should be taken to ensure research generates a dataset of a size which can be handled by the subsequent analytics programs."*. This is an important aspect to consider, as the volume of data produced can be on a very large scale. This could break the confines of some analytical programs'

constraints. Other Big Data characteristics (namely: variety, veracity, velocity and virtue) and the type of techniques applied by the researcher can have an influence on the choice of analytical tool adopted to achieve their aim(s). The naming of this section has been selected to align with its focus on preparing the data for the analysis, helping to identify whether the chosen analytical tools need to be changed to handle the dataset(s) in question and to establish which techniques (in our case, change point identification, sentiment analysis and machine learning) should be applied to analyse the data to assist in responding to a research aim and answering relevant research questions. The selection of techniques to analyse the data is a complex process that is dependent on the investigators' level of experience of the techniques in question while also ensuring that they will suit the dataset(s) chosen. For example, the selection of sentiment analysis techniques for a newcomer to a developing field can be fraught with difficulties as different papers suggest different techniques to use and most do not provide a concrete path to understanding the basics before choosing what path to follow. Social media analysis is a developing area and at present one does wonder if the techniques available are effective enough for any given specific domain, whereas in other fields techniques may well have been tried and tested over many years. In our experience within the pilot study in section 4, this led to it taking a considerable length of time to make a decision, which is why it's appropriate for this consideration to have a stage of its own. Another consideration to make at this stage is whether the researcher has the appropriate equipment to process Big Data and explore the intricacies of the dataset chosen using the desired tools. For example, initially within our research, using the R language presented some issues when processing a large amount of data, as R Studio is single threaded. This meant the PC being used was inadequate and required an upgrade due to poor single threading performance. An assessment must be made early on as to whether the PC or Cloud selection has the processing power to analyse the data in a reasonable amount of time (or indeed at all if there are memory considerations).

6) Stage 6 was originally entitled “Implementation/ Publication/Action” and has been renamed to “Implementation” here. The GSR approach originally emphasised that social media research is in its infant stages and that the likelihood is that the work being carried out will be exploratory. Any outcome *“successful or otherwise should be communicated”* to the interested communities to build on this in future work, which is the same in business as in research. To assist in these steps the researcher can include the good practice from the SMDS approach on “publishing” to “reuse/sharing” and “preservation”. Publication is one of the steps in this section as dissemination of research is clearly vital. The GSR lifecycle emphasises successful outcomes, but as this is now named "Implementation", there is a new focus, more appropriate for research, on making sure the project requirements and specifications as previously outlined above are implemented in practice so as to achieve the aims of the project. For example, in this step we extracted the data with NVivo NCapture, cleaned them and analysed them to detect the sentiment

within each Tweet and identify significant changes of sentiment within the timeframe over which the data were collected by using R. It was appropriate that this all took place within this phase, as one step flowed to the next with purpose and direction to contribute to the aim of the project. In addition, to this, ethical consideration must be given further thought at this phase to how any data are shared and preserved, but this data management process will not be discussed in this paper, as we shall focus on the legal and ethical considerations of social media data usage, which will look in particular at publication dilemmas. Publication is included in the last phase of the lifecycle instead as we must implement and (in particular) evaluate *before* we can publish within the research context. In our own context, had we attempted to include publication here alongside analysis, this stage would have become confused by the lack of evaluation. Furthermore, given the paucity of the quality of social media data, we required additional focus on relevant cleaning of the data and attempting to consider publishing at the same time would have resulted in a loss of momentum.

7) Stage 7 (Evaluation) is included in the lifecycle due to the immaturity of social media research compared with other more established research fields. There is a focus on the evaluation of exploring what value there is in social media research compared to traditional methods. It suggests that this stage will confirm whether not social media was specifically required "to respond to a business or citizen need". This stage will remain the same as outlined in GSR's lifecycle but with a rather different focus. Where the GSR strategy considers whether or not there was value in the use of social media data, the researcher's focus will be on how effective the use of such data was in addressing the research aims and questions. A stage devoted to evaluation is important, as through evaluation we can identify whether our techniques have been effective in answering any research questions. For example, in our case, we aim to consider whether using a lexicon dictionary approach over machine learning for detecting sentiment provides a greater level of accuracy within the framework we have set.

8) Stage 8 has been renamed from "Business as Usual", as it is in the GSR lifecycle, to "Knowledge Management" in order to fit the research context. The original purpose of this phase remains, but with the addition of publication to emphasise its importance in this context. This phase re-evaluates research techniques in order keep research up-to-date with any modern research techniques and to think how about how any knowledge gained about social media research methods themselves can be transferred to others to instil good practice. This stage can be commenced once a significant part of the cycle is completed. Publications are crucial way of sharing good practice within the research community and can then lead to subsequent further research after interactions with the community, leading us back to stage 1 to begin a new project and frame suitable new research questions. The pilot study's outcome in section 4 has informed us that this original lifecycle with a series of changes

can be placed into a research context that is effective in guiding social media projects.

It is important to note the lifecycle is not only to be used as a single iteration. A researcher can repeat stages to develop the project through one or many iterations. Moreover, this lifecycle itself will be further evaluated when cycling through it again in the implementation phase with four new datasets. Background on these additional datasets chosen will be discussed in section 3.1.3.2 once the step 1 rationale and step 2 research methodology are outlined in the next chapter that will be implemented in this project.

3.1.1 Step 1: Rationale

There is a business need to find datasets that allow for real-time monitoring of potential risks (GSR, 2016). Social media is relatively a new area of analysis that can be quicker and cheaper than some other different methods of analysis and data (such as NHS data) is available nearer to real time.

As previously outlined in section 1.3, this project looks to enhance the analysis of social media in the context of public (dis)order events. Additionally, we will investigate how social media data are stored (big data issues), collected (data extraction), analysed (text mining and sentiment analysis) and then disseminated (to the police to help predict when disorder may occur). We are focusing on applying sentiment analysis techniques, and then analysing these data through different time series techniques to identify key elements in demonstration together with the creation of a model to analyse social media data to predict the escalation of such events.

In the next section, we will select a research methodology and explore the chosen approach for this research project in step 2.

3.1.2 Step 2: Selection of Research Methodology

As previously outlined in section 3.1, the project will use a case study methodology as framework for the social media research. In order to understand what a case study means in research, we will discuss the theory of the case study and its application to a social media context.

Yin (2015, p46) outlines *"The essence of a case study, the central tendency among all types of case study, is that it tries to illuminate a decision or set of decisions: why they were taken, how they were implemented, and with what result. (Schramm, 1971, emphasis added)"*

This definition above according to Yin (2015) cites 'decisions' as the main area of focus in a case study. *"Other common cases include "individuals," "organizations," "processes," "programs," "neighborhoods," "institutions," and even "events.""* (Yin, 2015, p46). For case studies, these five components of a research design are especially important.

According to Yin's (2015) first point, the scope of the case study is to investigate the "contemporary phenomenon in depth" that is encompassed within a "real-life context", as its highly pertinent to the phenomenon of this PhD study. The data used in this thesis are derived from four Twitter-based case studies. There are many research papers (Ahmed, Bath, Demartini, 2017; He, Zha & Li, 2013; Lin, Hoffman & Borengasser, 2013) that have adopted a case study approach. Even though the datasets have different aims and study various event/ phenomena, it may be possible to identify trends, patterns and principles that work across social media research and numerous potential studies. The area of interest in our case is to analyse social media data using sentiment analysis in the relation to public order.

The comprehensive qualitative accounts produced in case studies help to explore, describe and explain the complexities of real-life situations that might not be captured in experimental or survey research (Dul & Hak, 2007; Woodside, 2010; Yin & Campbell, 2018). This strategy centres its data collection on historical datasets or archive documents. This allows for exploratory and explanatory/ descriptive analysis of changes pursued over a period of time. Case studies rely on archival data, so there is a need to be aware of all the possible historical biases and to proceed with caution in interpretation of results and findings. Within mass media, it may be helpful to choose two different media that are believed to exhibit opposing views, so a more balanced picture might emerge (Dul & Hak, 2007; Woodside, 2010; Yin & Campbell, 2018). These additional findings and use of other sources would help further widen the view to get a complete version of events. In this process, it is important to be conscious of the benefits and limitations to the application of this case study to be factored into the research approach (Dul & Hak, 2007; Woodside, 2010; Yin & Campbell, 2018). According to Dul & Hak (2007), Woodside (2010) and Yin & Campbell (2018) these comprise: -

Benefits:

- A case study can help simplify complex concepts.
- Case studies expose the participants to real life situations which otherwise is difficult.
- Case studies collect greater detail that may be more difficult to obtain using other research designs. The data tend to be richer and has greater depth than obtained through using experimental designs.

- The variations of intrinsic value, collective and instrumental approaches to case studies allow for both qualitative and quantitative analyses of the data.

Limitations:

- It may be difficult to locate a suitable case study for all subjects.
- Case studies contain the study of observations and perceptions of a single person. There is a chance the person presenting the case study might not be aware of information that are pertinent to the study. Additionally, a problem arises in validation of the solutions, as there may be more than one way to view the data.
- One of the main criticisms of the data gathered is that it cannot always readily be generalised to the wider population.

The case study's benefits provide for experiential learning, but one must consider the drawbacks to minimise the bias in the outcomes (Yin, 2015). The second point outlined by Yin (2015) emphasises that "phenomenon and context" are not always clear in "real-life situations". Yin (2015) follows on by saying "phenomenon and context" may not be "sharply distinguishable", therefore, other relevant methodological characteristics become involved through design and data collection features as follows: -

- "copes with the technically distinctive situation in which there will be many more variables of interest than data points, and as one result"
- "relies on multiple sources of evidence, with data needing to converge in a triangulating fashion, and as another result"
- "benefits from the prior development of theoretical propositions to guide data collection and analysis."

Among the variations in case studies, a case study can include single or multiple cases, can be limited to quantitative evidence, and can be a useful method in performing an evaluation. Dul & Hak (2007), Woodside (2010) and Yin (2015) shows that case study research includes all the qualities of an "encompassing method", as it covers the logic of design, data collections techniques and approaches to data analysis, so it does not use one design feature alone. Case studies are effective when a "how" and "why" question is being asked about a set of contemporary events (Yin, 2015) with no control over behavioural events and the focus of study is on a particular phenomenon. In the case of this project, it includes a recent UK-based social movement and interactions between the public and police in a series of demonstrations (Ahmed, Bath, Demartini, 2017; He, Zha & Li, 2013; Lin, Hoffman & Borengasser, 2013). This approach preserves the connection between the context and its phenomenon (Dul & Hak, 2007; Woodside, 2010), which retains the capacity to deal with the complexity of the case studies. Each case study chosen represents prototypical features of the types of data collection

scenarios that are pertinent to the researcher's experience. For example, these dimensions include varied levels of political contention and timescales of a long/ short duration and of media such as text, images and videos (Dul & Hak, 2007; Woodside, 2010; Yin & Campbell, 2018). The range of these case studies provide for triangulation of different contexts of social movements and interactions with police when examining short time intervals of social media data within, across and between cases.

There are various types of case studies that provide differing perspectives present in case analysis which informs research in contexts that may correspond with one or more prototypical dimension from one of the chosen case studies (Yin & Campbell, 2018). In between case analysis, this may provide information about the level of impact on the various prototypical dimensions within the potential short time span of the social media datasets (Ahmed, Bath, Demartini, 2017; He, Zha & Li, 2013; Lin, Hoffman & Borengasser, 2013). The case analysis may enable one to generalise across Twitter datasets, for example, in relation to social movement and political contention. These four chosen datasets as specified in stage 3 of the social media lifecycle are socio-economic and political in nature, but each case might exhibit a varied level of ephemerality. As a result, the analysis will be focused in a specific area, so this may be less useful in the context of different types of events, such as concerts or football events. However, this is not of a concern as we are primarily focused on demonstrations relating to public order (Ahmed, Bath & Demartini, 2017; He, Zha & Li, 2013; Lin, Hoffman & Borengasser, 2013). Furthermore, the combination of these four case studies enables a wider general understanding of ephemerality throughout Twitter and possibly other social media websites.

The following parts of the study will describe the case study method adopted for the research that will sit within the social media strategy outlined in section 3.1, in which application of data collection and analysis techniques are discussed. A description of each case study will be outlined in the section 3.1.3, and the associated data collection procedures in section 3.1.4 and proposed analyses are described in section 3.1.5. A list of query terms for each case study are outlined in section 3.1.4.2.

3.1.3 Step 3: Data

The following sections will outline the purpose of the pilot study, using its results to narrow the focus of the project, to provide information about each event, and then provide reasons for the datasets chosen for study.

3.1.3.1 Datasets

The dataset used for the pilot study (discussed in section 4) is based on the 2015 Baltimore riots in the USA. This data had previously been collected by the researcher

using Nvivo software so was utilised again within the pilot study. The pilot study helps to inform future collected datasets, in which previous text mining and data mining scripts can be rebuilt and re-run on the new datasets.

The Baltimore riots occurred between 18th and 25th April 2015. The main cause for the riots was due to the arrest of the black American called Freddie Gray on 12th April for possession of a 'switchblade', but court documents said this was a false accusation (Baltimore City, 2019; Ortiz, 2015; Woods & Pankhania, 2016). Gray sustained injuries during the arrest and subsequently in the police van through police brutality. Gray arrived at the police precinct after a series of stops and was then found to be unconscious. A paramedic was called and Gray was taken to hospital in a coma. After Gray's subsequent coma, protests began outside the Western District police station. They started on 18th April and were relatively peaceful (Baltimore City, 2019; Ortiz, 2015; Woods & Pankhania, 2016). Gray underwent two surgeries on 19th April to attempt to save his life, but the operations failed and Gray died (Baltimore City, 2019; Ortiz, 2015; Woods & Pankhania, 2016). Protests continued through to 24th April, but this escalated on 25th April and the peaceful protest turned violent, with riots erupting. A curfew was placed on Baltimore on 28th April, which resulted in peace. Legal proceedings for charges against specific protesters and police continued. Subsequently, on 1st May, six police officers were charged in connection with Gray's death (Baltimore City, 2019; Ortiz, 2015; Woods & Pankhania, 2016).

The results from the pilot study showed that the American English used in these tweets differed from standard British English. Moreover, the American police's strategic approach was shown to differ from that of the British police. As a result, this impacted the type of language used on social media. It was decided that the differences were too great in comparison to the British context and thus that future datasets would be based on UK public order events which largely comprise demonstrations (organised events in-line with the police), rather than protests (a protest occurs on demand due to civil unrest being caused by some level of injustice) (HMIC, 2009). The decision was thus made to focus on demonstrations, which will permit the establishment of a sound model for the data analysis.

3.1.3.2 Chosen Datasets

3.1.3.2.1 2015 Million Man March (MMM), London

The Million Mask March (MMM) is annually held in London (includes other countries, such as the US) on November 5th every year since 2012 except during COVID-19 pandemic, demonstrations have been organised by a hacktivist (someone who enters a computer system without permission to achieve a political aim (Cambridge Dictionary, 2020)) group called "Anonymous" that is part of a larger demonstration worldwide. The motive for each march varies but does include some consistent themes including corruption in politics and self-governance (Johnston & Gayle, 2015; Sims, 2016).

The “Anonymous” organisation and its representatives demonstrate against capitalism in London between 18:00 to 21:00 starting in Trafalgar Square (Johnston & Gayle, 2017; Turner & Finnigan, 2015). In the UK on this day Guy Fawkes is celebrated with a bonfire night that contains fireworks. The 2015 MMM held in London focused on proposals to increase powers of the security services, as the government published a bill seen to revive the controversial snooper’s charter the day before the demonstration (Johnston & Gayle, 2017; Turner & Finnigan, 2015). Facebook indicated that around 20,000 people planned to attend the demonstration wearing Guy Fawkes masks in an effort to recreate the closing scenes of the cult movie V For Vendetta (Johnston & Gayle, 2017; Turner & Finnigan, 2015). The actual numbers of demonstrators that attended the event were more than 1,000, but lesser than outlined on Facebook (Johnston & Gayle, 2017; Turner & Finnigan, 2015).

The march started at 18:00 from Trafalgar Square, and finished at Parliament Street, Whitehall, where the demonstration lasted till 22:45, even though police reminded demonstrators’ that the curfew is at 21:00 (Johnston & Gayle, 2017; Turner & Finnigan, 2015). The event started peacefully, but in time demonstrators began to clash with the police. Some demonstrators showed glimpses of criminal damage, using drugs and use of offensive weapons (Johnston & Gayle, 2017; Turner & Finnigan, 2015). For instance, a police vehicle was vandalised and set alight, and demonstrators went outside of the agreed route to Buckingham palace throwing cones and fireworks at police horses (BBC News, 2015; Johnston & Gayle, 2017; Turner & Finnigan, 2015). Furthermore, allegedly Terry Small was unlawfully hit by a police baton and an Aston Martin car collided into a demonstrator. The driver drove off at speed before the demonstrators descended on the vehicle (Johnston & Gayle, 2017; Turner & Finnigan, 2015).

The march finished at Trafalgar Square, where police made a containment (also in occurred in Parliament Street, Whitehall) in the location to curtail the protest (Johnston & Gayle, 2017; Turner & Finnigan, 2015). There was a heavy police presence of around 2,000 officers at the march with riot vans to prevent potential unrest that was expected at the event (Johnston & Gayle, 2017; Turner & Finnigan, 2015). The police arrested 50 demonstrators and 3 police officers along with some demonstrators had to be treated at a hospital after being injured (BBC News, 2015; Gayle & Johnston, 2015). Additionally, the police horses had to be treated by a vet for injuries sustained (Turner & Finnigan, 2015).

3.1.3.2.2 2016 Million Man March (MMM), London

In 2016, the MMM march focused on the *“government's disregard for migrants, for the poor, the elderly and the Disabled, we have seen the capital, profit and greed of the few put before the well-being of the many and we say enough is enough”* (The Guardian, 2016) This march had strict conditions imposed on it by the police due to

last year's march disorder, as outlined in 6.1.1. On Facebook nearly 20,000 people had indicated to attending the event page and warned *"the police are not your friends"* (The Guardian, 2016).

The march started off peacefully until about 1,000 demonstrators headed towards Whitehall, where tempers flared where police had formed a ring of steel outside parliament (The Guardian, 2016). Static demonstrations are permitted to occur at Trafalgar Square, Richmond Terrace and Parliament Square (Nagesh, 2016). There were exchanges between police and demonstrators, who were reminded to stay in boundaries of agreed route and argued to be allowed past (The Guardian, 2016). Small numbers of flares and fireworks were lit outside of Westminster Abbey (BBC News, 2016; The Guardian, 2016). At 19:00 the police reportedly arrested 10 protesters. The number of protesters at Parliament Square reduced to hundred at 19:30, when a man was being led away by police (The Guardian, 2016). Just before 21:00 there were chaotic moments, as riot police made an arrest and a group of demonstrators swore at the police and threw glass bottles (The Guardian, 2016). At 21:00 the police reported that the number of arrests rose to 33 (Sims, 2016). The police by 22.45 had made 47 arrests, (The Guardian, 2016) mainly based on drug offences and obstruction of officers, but according to BBC (2016) 53 were arrested. The MET outlined there were pockets of disorder, but most protesters in the march were peaceful, therefore, no containment was required (BBC News, 2016).

The social media data analysis research on MMM is limited. When limiting the search for MMM 2015 and MMM 2016 events in the UK no research papers are found. However, there are some papers that focus on MMM only in a wider view (Armstrong, 2017; Knight, 2018; Harbisher, 2016) than specific to UK.

3.1.3.2.3 2016 Dover Demonstration

The Dover demonstration occurred on 31st January 2016, where far-right groups (including The National Front, SouthEast Alliance, NorthWest Infidels, and the East Kent English Patriots) and left-wing anti-fascist groups (Kent Network Against Racism and Dover Stand up to Racism) demonstrated in the port of Dover over the UK's position on immigration policy (Gayle, 2016; Lennon, 2017; Osborne, 2016). On the same day, there was a parallel demonstration by both sides in Dewsbury in West Yorkshire on the same issue.

The march was staged by the 'Kent Anti-fascist' network and different left-wing groups were held at the Market Square in Dover's town centre at 11am (Gayle, 2016; Lennon, 2017; Osborne, 2016). The far-right demonstration began at Dover Priory railway station at 2pm. Tensions ran high throughout the event, where a strong police presence was on display. As the time neared the start, a group of masked anti-fascists

broke off from the Market Square at 12.30pm towards the train station where far-right groups were gathered (Gayle, 2016; Lennon, 2017; Osborne, 2016).
Police arrested many demonstrators when the far-right 'East Kent Alliance' clashed violently with their rival group 'Kent Anti-Racism' network (Gayle, 2016; Lennon, 2017; Osborne, 2016). Nine people were arrested, some for the possession of offensive weapons, breaching the peace, violent disorder and a range of public order offences.

The research Twitter data analysis on the 2016 UK Dover demonstration could not be found when searching for scholarly publications. There was research based on immigration policies (Bartlett & Norrie, 2015; Jensen, 2016), which is on the use of social media data.

3.1.3.2.4 2016 Anti-Austerity Demonstration

The Anti-Austerity March took place in London on the 16th of April 2016 and was organised by the People's Assembly Against Austerity and other unions, such as Unite (BBC, 2016; Broomfield, 2016). Some of the unions and groups attended included, the Campaign for Nuclear Disarmament, the National Union of Students, the National Union of Teachers and Stop the War Coalition (ITV News, 2016). The march was against the Conservative budget cuts to budgets of health, homes, jobs and education (BBC, 2016; Broomfield, 2016). The demonstrators called for David Cameron's resignation because of the link to his father's offshore company leaked in the Panama papers (BBC, 2016; Broomfield, 2016).

The march began at 1pm on Gower Street near the University of Central London. The demonstrators made their way through the streets for a rally in Trafalgar Square (1.1 miles in distance), where the demonstration lasted till 6pm (BBC News, 2016; Grierson, 2016; ITV News, 2016). There were 150,000 people reported to be involved in the rally in London against the cuts (ITV News, 2016). Labour Shadow Chancellor, John McDonell addressed the crowd to call for David Cameron to resign as Prime Minister of the Conservative party (BBC News, 2016; Grierson, 2016; ITV News, 2016). Unite Union leader, Len McCluskey made a reference to the Panama tax haven scandal that involved the prime minister. Additionally, Green Party leader, Natalie Bennet called for the Tories to be out and not just David Cameron (BBC News, 2016; Grierson, 2016; ITV News, 2016).

The research social media analysis on Anti-Austerity coverage is on a low scale, but when focusing down on this specific event in the UK this could not be found. The research on Twitter data around this event focuses on other countries than the UK (Bailo, Vromen, 2017; Barisione & Michailidou, 2017; Gerbaudo, 2017; Karyotis & Rüdig, 2018; Theocharis et al., 2014).

3.1.3.3 Justification for the Datasets

There are many reasons behind the choice of datasets that will be analysed in the thesis: -

- Public order events tend to be mostly demonstrations in the UK rather than protests (Rogers, 2011). Each of the demonstrations has been chosen based on its level of presence and discussion through online media (social media, news and blogs). In addition, these demonstrations have been picked as past events, such as 2010 Tuition March and 2011 London riots and have already been thoroughly researched (Cammaerts, 2013; Fuchs, 2012; Procter, Vis & Voss, 2013; Theocharis et al., 2014; Vis, 2013) and analysed. Proctor, Vis & Voss (2013) good practice to use search terms to identify relevant keywords to remove irrelevant tweets will be applied to the project to create a list of keywords for public order events to identify relevant/irrelevant tweets. The outcomes of more recent demonstrations may elicit new information that can inform the project. In the most recent demonstrations in the UK, the police are using the most up-to-date strategic approach to increase public safety. These methods are outlined in section 2.2.2.

- Over the course of 2015 to 2016, observations were made that these four chosen demonstrations in the UK have received high levels of discussion on social media, news and television, whereas other demonstrations were not as visible. This means the data can be abstracted, generalised and evaluated to develop understanding to make inferences from the data to drive actions in new contexts. These four demonstrations can be compared with respect to the use of language in related tweets and whether any significant occurrences were sparked over time. This may help to uncover important information that could be considered beneficial to increasing public safety at future UK demonstrations.

- Violent acts were recorded at three of the chosen demonstrations (Johnston & Gayle 2015; Lennon, 2017; Sims, 2016). These demonstrations will be investigated to consider how the violent incidents were prevented from escalating to a full-scale riot. The Anti-Austerity March was the only peaceful event (BBC, 2016) so can be used to make comparisons with the other demonstrations in relation to timescale and season. As the MMM has been an annual event since 2012, this can be compared with previous and future MMMs.

Having outlined a possible lifecycle for social media research, selected a research methodology and identified suitable data, in section 3.1.4 there is a discussion of the

ethical and legal considerations that are made throughout the social media research lifecycle.

3.1.4 Step 4: Tools and Output

The datasets have been chosen, and in this step, we will explore which social media platform(s) will be chosen to extract data related to the datasets along with tools for extraction.

3.1.4.1 Single or Multiple Platforms for Extraction

Single and multiple platforms have been explored from which to extract data, but if many are used it would be difficult to integrate the data, as the metadata would different. Additionally, it would be difficult to address the user population, as it can be difficult to make a comparison between platforms (Sloan & Quan-Haase, 2016). To extract and analyse data from multiple platforms, alongside managing and organising it, requires resources that are not available in a study of this scale. Therefore, it has been decided to focus solely on Twitter. Twitter is the most open platform and is widely used by the research community (Sloan & Quan-Haase, 2016). Twitter can be seen to have advantages over its competitors due to the short character limit per tweet that encourages Twitter users to provide live updates, at any time, in any location and on any device (Sloan & Quan-Haase, 2016). Since 2017 the character limit has 256 characters (before it was 140 characters) per post based on the Short Message Service (SMS) which keeps the message concise, easy to read and write. In relation to a topic of interest, the retweet capability helps to further disseminate a message from various movements that occur, increasing awareness of a given demonstration. The hashtag allows users to spontaneously participate and connect on shared topics in the public arena. Live tweeting can help connect people offline and online, where online communication makes the discussion accessible to people who are not physically present at the demonstrations. This can provide a public record of the events' activities, such as cancelled meetings and people that cannot attend the gathering (Sloan & Quan-Haase, 2016).

Some social media platforms, such as Facebook, have more complex privacy controls than Twitter. For example, (Croeser & Highfield, 2014) Facebook posts are less accessible by the public, as the user has greater control over the privacy settings. The hashtag functionality is less visible than Twitter, as the Facebook user has the option to post a message as public, friends of friends, or only me, whereas on Twitter there is a binary choice between public and private (Sloan & Quan-Haase, 2016). As a result, Facebook posts are more difficult to obtain than Tweets, as open authorisation is required from the user. If the Facebook posts with hashtags or groups are set to public, then that data can be extracted via Facebook's Graph API via a registered user

account. Even though Twitter is the platform of choice and no data will be extracted from any other platform.

Once a platform has been chosen, the next step is to prepare for the data collection from selected platform(s). According to Sloan et al (2016, pg110), the following questions must be considered when choosing a tool: -

1) *"What are my main criteria for selecting data from this platform? (Basic approaches for collecting data from social media)"*
2) *"How much data do I need? (Big vs. small data)"*
3) *"What is (unproportionally) excluded if I collect data this way? (Collection bias)"*

In answer to question 1, it is important to consider the time duration, as the data collected for different periods (hours, months or years) of time can significantly impact the statistical outcome (Sloan & Quan-Haase, 2016). In addition, other criteria include those based on topics and keywords, metadata and based on user accounts. The focus of interest is on the topic rather than user accounts and the other highlighted issues, which is further emphasised in sections 3.1.4.1 and 3.1.4.2. With respect to question 2, big data has become prominent, especially in the context of social media data, as the rate of content shared, and growth of the user-base, is increasing. There are examples (Kwak et al, 2010; Schroeder, 2014) that demonstrate collected data from a network of millions of users. With big data, there are questions of data storage, processing infrastructure, limitations of the API and ethics (refer to section 3.1). The size of the dataset is not the main focus, it is how the data are both composed and collected as emphasised in question 1. This has been addressed by choosing most relevant terms is extract the demonstration data in section 3.1.4.2. As few guidelines exist on the level of permitted data collection, the researcher has to decide on the amount of data to be gathered (Sloan & Quan-Haase, 2016). Sloan et al (2016) suggests up to 10% of social media data collected is useful. The quality can depend on the collection criteria as outlined in section 3.1.4.2. Considering question 3, there can be problems (Bruns & Stieglitz, 2014; Ruths & Pfeffer, 2014) in the collection of data, which can bring a specific bias to the dataset. The most commonly used approaches include biased social media populations, non-transparent access restrictions to user data (Morstatter et al, 2013) and sampling biases, such as a focus on collecting tweets based on users sharing their geo-location. This sub-group might not be representative of every user on Twitter because they do not all share this information. Based on this acquired knowledge, the selection process outlined in both sections 3.1.4.1 and 3.1.4.2 outlines technical ways to extract data based on topics/keywords that can help effectively address question 1.

Social media data can be collected with the use of automated tools to collect, clean, store and analyse the high velocity of large volumes of social media data, which can be retrieved near real-time in some instances, is ongoing. The characteristics of such data differentiate themselves from materials created from traditional research methods,

such as surveys and ethnography. There are many options to collect social media data, including the purchase of data from an authorised re-seller, such as GNIP or SIFTER or the use of other non-paid tools that can extract data (such as TAGS or COSMOS) or a programmer can extract data from an API. APIs are a set of building blocks, such as tools, protocols and information, where aspects can be re-used to permit the building of programs. Twitter platform chosen has a large user-base, detailed documentation is provided on how their APIs work and there is a wide developer community (Batrinca & Treleaven, 2014; Vis, 2013). The section of 3.1.4.1 for chosen platforms and a list of chosen tool(s) (Table 2) have been carefully considered to help (refer to section 3.1.5) ensure a more complete dataset and avoid bias in the analysis.

[Intentionally Left Blank]

Tool	Platforms	Data Storage	Text Analysis	Visualisation	Cost	Entry level	System requirement
DiscoverText **Sifter (part of DiscoverText)**	Twitter, Google+, Text Files, Email, Blogs, and open-ended answers on surveys, Facebook	Unknown, but will be restricted as on their server.	Yes	Yes	Reduced cost for university staff & students	Beginner/Intermediate	Web accessible
GNIP	Twitter, YouTube, FaceBook, Instagram, Google+, Reddit, etc.	Unknown	Yes	Yes	Quotation required	Unknown	Web accessible
Exploratory.io **(R language)**	Twitter	Unknown	Yes	Yes	Free or paid version	Beginner	Windows/ Mac
Google Sheets add-on: **Tags**	Twitter	Limited to 10MB per each sheet	No, unless using Google add-ons No, unless using Google add-ons	Yes	Free	Beginner	Web accessible
Google Sheets add-on: **Twitter Archiver**	Twitter			No, unless using Google add-ons	Free (one extraction only), add-ons cost	Beginner	Web accessible
Google Sheets add-on: **Blockspring**	Twitter, Linkedin, Recruitment websites, FaceBook, YouTube News sources, Blogs,		Yes	Yes	$10 per month	Beginner	Web accessible
Morph.io	Web scrape's website, alternative is using Google Sheet's web scraping add-ons for ease of use.	Unknown, but will be restricted	No	No	Free	Intermediate	Web accessible
NVivo with add-on tool for browser: Ncapture	Multiple sources, such as Twitter, FaceBook, WordPress, Blogs & News.	PC capacity	Yes, NCapture can capture tweets manually, but data is only accessible through NVivo's other software application. Data can be exported		Free for students, otherwise it costs	Beginner	Windows/ Mac
COSMOS	Twitter	PC capacity	Yes	Yes	Free for universities	Beginner/Intermediate	Windows/ Mac
Twitter Capture Analysis Tool	Twitter	Unknown	Yes	Yes	Free	Unknown	Windows/ Mac
Mozdeh	Twitter	PC capacity	Yes	Yes	Free	Intermediate	Windows/ Mac
Apache Spark/ Apache NiFi		Dependent on the package bought.	Yes, but other parts of the Hadoop eco-system are required		Free	Advanced	Web accessible/ Windows/ Mac
Using R and installing TwitteR package	Twitter	PC or server capacity	Yes, has to function with other packages	Yes, but with other R packages	Free	Advanced	Web accessible/ Windows/ Mac Help to get started: https://nbviewer.jupyter.org/
Using Python and its packages: use pip and install python-twitter	Twitter	PC or server capacity	Yes	Yes	Free	Advanced	

Table 2 Data Acquisition Tools

The tools preferred from Table 2 are discussed below: -

- The purchase of historical Twitter data can be costly. If the tweets of interest are not extracted for free from Twitter in the seven-day period from today's date, then the Twitter data has to be bought from a licensed reseller. The historical data from Twitter can range from inexpensive to very expensive as it depends on both query type and time of retrieval (Sloan & Quan-Haase, 2016). Estimates of cost can be directly generated from some tools, such as SIFTER, whereas GNIP requires direct contact to obtain a quotation which may take longer to process (Burgess & Bruns, 2012). SIFTER is a service that provides search and retrieve access to Twitter's undeleted tweets via GNIP's historical PowerTrack. This tool along with GNIP and DiscoverText can be the least inexpensive method to extract Twitter's historical data. If data is bought from SIFTER, this has to be used in DiscoverText to access, analyse and export the data. DiscoverText provides other packages to extract, transform and load the data for analysis. DiscoverText provides a free trial, but if data are bought via SIFTER, then the enterprise package is available for up to 60 days, but after this it costs $750 per month for students. Additionally, there is a basic or professional package, which is cheaper (DiscoverText, 2018).

- Google Sheets is an accessible tool "*for data scientists interesting in manipulating or engineering data, and does a great job of making the data easily visible as it is edited"* (Slater, Joksimović, Kovanovic, Baker & Gasevic, 2016, pg4). This tool is *"not useful for engineering variables in extremely large data sets, around one million rows and above, but they are excellent tools for smaller-scale feature engineering, and for prototyping new variables in subsets of a much larger data set."* (Slater, Joksimović, Kovanovic, Baker & Gasevic, 2016, pg4). Google Sheets is useful for processing smaller subsets of large datasets. TAGS (https://tags.hawksey.info/), a Google Sheets add-on, enables a user to extract Twitter data and directly archive it into a spreadsheet. Search terms must be entered into the sheet where specified. You can run automated CRON (a time-based job scheduler) job to collect the Twitter data by the hour or day.

- Apache Spark is an in-memory data processing engine that has a *"development API to allow data workers to efficiently execute streaming, machine learning or SQL workloads"* (Karim, 2017, pg180), providing fast iterative access to datasets. "Apache NiFi supports powerful and scalable directed graphs of data routing, transformation, and system mediation logic." (Apache NiFi, 2019). Apache NiFi is a data flow tool that can automate the flow of data from any source and then distribute that to various systems (Hadoop and Spark) to gain insight from the data (Dasgupta, 2018). Both Apache NiFi and Spark can store data in MongoDB (NoSQL database) or Hadoop Distributed File System (HDFS).

- Exploratory Desktop (https://exploratory.io/) provides an interactive and reproducible real-time data wrangling and analysis experience powered by R (https://www.r-project.org/). Exploratory provides an overview of data through its default visualisations and it abstracts away the user from writing R code, but the user can export the R code to utilise in other applications. This tool can extract tweets directly from Twitter. The disadvantage is the manual retrieval of the data rather being an automated process without user intervention. Importing data is relatively straightforward process compared to COSMOS (http://socialdatalab.net/COSMOS) and Mozdeh (http://mozdeh.wlv.ac.uk/), though it requires a more specific format that may be challenging to a beginner user.

The tools Table 1 have a purpose that come with a set of benefits and limitations. The chosen applications have been selected based on ease of use, storage capacity, cost and automation. Most tools require a user to manually extract tweets (unless a server is available), which can be difficult if an event last hours or days. If the user cannot be there, tweets in that time will not be collected so, therefore, the data gathered may not provide an accurate reflection of the event. The data collection tools do not extract all the metadata from Twitter's API. The level of incompleteness in the dataset can vary, meaning some lines of enquiry may not be explored, which can impact the results of any analysis.

Apache NiFi, Apache Spark and TAGS can automatically extract tweets from Twitter without the user being present, which means data are collected with consistency in a period of time. Within this research Apache NiFi and Apache Spark are preferred choice over TAGS, as they can store a larger number of tweets, whereas Google spreadsheets has limits of 10MB (Dasgupta, 2018). This means that new spreadsheet documents must be manually created every time to keep on continuously collecting data. The likelihood of duplicated data is high, as it can be difficult to know when to extract the data again after the initial retrieval. Not all data can be captured in a time block depending on volume and Twitter API restrictions (Twitter, 2019). An optimal method is to create a server and stream the data in with the use of Python, R, Apache NiFi or Apache Spark. Data can be stored within Hadoop and/or a database (Dasgupta, 2018). Otherwise, the data must be bought from Sifter or GNIP. In this project, SIFTER has been used to buy historical data, as past events are required. In addition to this, data collected from other platforms lack extensive metadata that can be retrieved from a tweet, whereas SIFTER (https://sifter.texifter.com/) can extract all metadata for every tweet.

3.1.4.2 Topics and keywords for Data Extraction

As previously outlined in section 3.1.4.1, the tool to be used to extract the data will be DiscoverText. To understand the boundaries of what is possible from the data collection is important, as the data may be limited and contain potential bias that can impact the analysis viewpoint (O'Neil & Schutt, 2014). The decision has been made to follow Sloan et al (2016) suggestion to collect data on the four chosen demonstrations based on keywords/hashtags to extract the most relevant data for the event. The keywords/hashtags were chosen on the basis of their relevance and popularity in relation to each of the demonstrations. As the events were long passed by the time the data were collected, it was made easier to identify the hashtags in question because of news, social media and online tools, such as "hashtagifme". Social media and news media clarify which hashtags to follow in the articles outlined in sections 3.1.3.2 and 3.1.3.2 and organisers of the demonstrations and the tweets on Twitter showed a common pattern in the hashtags used. These hashtags were checked with "hashtagifme" on their popularity and other associated words based on their strength to ensure the most relevant and highly used hashtags are chosen. Keywords were searched for all tweets, but this could present more irrelevant tweets as it could be based any topic. Therefore, hashtags may be more appropriate, and users tend to tweet with a tag for a specific topic, which might increase the relevance of the data collected for each demonstration event. The most prominent and relevant hashtags for the events we selected (as noted in section 3.1.3) are presented below:

- **Dover protest 30th January 2016**
    - #Dover and #antifa
    - 27/01/16 - 01/02/16
    - Number of days: 6
- **Anti-Austerity march - 16th April 2016**
    - #4Demands and #London
    - 13/04/16 - 18/04/16
    - Number of days: 5
- **Million Mask March - 6th November 2015**
    - #MillionMaskMarch, #Anonymous and #MMM2015
    - 03/11/15 - 08/11/15
    - Number of days: 5
- **Million Mask March - 6th November 2016**
    - #MillionMaskMarch, #Anonymous and #MMM2016
    - 03/11/16 - 08/11/16
    - Number of days: 5

The chosen hashtags were based on relevance, but had to limit number of hashtags chosen as more hashtags would have led to an increase in funds required, beyond those available. As a result, the number of hashtags chosen was restricted, so it was

vital to identify the strongest associated hashtag for each demonstration using "hashtagifme". The selection of the time frame was allocated based on the investigation of each hashtag. The timescale was determined by when the hashtag was first and last used in relation to the demonstration. The tweets were manually examined to identify the start and finish points in that time cycle, which resulted in the time frames specified for each demonstration. These tweets based on the hashtags were extracted without additional filters so as not to restrict anything (such as leaving a specific users or geo-coded data) from what is gathered. This was deliberately done as the data could not be examined in detail before collection in any great depth. Therefore, it was decided to retrieve all the data, and then assess it afterwards in terms of what to leave out.

3.1.5 Step 5: Analysis

We have already explored analysis techniques, but in this section the focus is the on analysis of data in a qualitative and quantitative way to effectively manage, organise and present the data to meet the needs of the research question. Additionally, it is vital to consider the representativeness of the data. As social media is prominent in society, it will generate large volumes of data. Research generating a dataset of this size must be handled by analytical programs as emphasised in section 3.1.

3.1.5.1 How to Analyse the Social Media Data

The quantitative and qualitative techniques that may be used in a project to analyse the volume and variety of social media data are outlined below:

Quantitative approaches

There are different techniques that can discriminate the data to gain insights from the frequency of discrete and categorical variables within social media datasets, which are (GSR, 2016; Sloan & Quan-Haase, 2016): -

- **Clustering –** This uses a series of algorithms to assign data to clusters, where they have similar characteristics, such as different types of topic (apples, oranges and bananas).

- **Classification –** Existing data collected can be compared with another dataset as correlations may be obtained (across time or another independent variable) to be used in a predictive capability. Models can be constructed to predict values (dependent variables) or categories.

- **Time series analysis –** The volume of data can be analysed on a specific user group, demographic or use of language based on keywords through a fixed time frame.

- **Geographic analysis –** The spatial element of social media data (geographic co-ordinates of a PC or mobile device) can be represented historically or near real-time to see the spread of an event such as a protest.

- **Relationship analysis –** This analysis focuses on interactions between users, where the number of fixed relationships is established to identify the links between each user and the number of responses to a post. This will help analyse the degree of engagement that occurs on a social media platform.

The data obtained by the research methods will undergo a form of text mining transformation, from text to numerical format that helps to quantify large amounts of data in simplistic powerful summary, which is known as descriptive statistics (Bryman, 2012). The summarisation of the analysis may include (GSR, 2016; Sloan & Quan-Haase, 2016):

- volume frequency – number of users, re-tweets and likes, volume per time slice

- textual semantics of keywords, comments, and hashtags, scores and rankings

- demographic data: name, gender and age, geographic location, and influences

We will focus on clustering, classification and time series analysis. The project aims to cluster keywords into categories and use time series analysis to predict what may happen. This may help the police adapt their public order strategy to keep the public safe. There are limitations with descriptive statistics as it reduces large amounts of data into a simple summary, which could risk distortion of the data or losing important detail that gives a fuller picture of the figures presented (Bryman, 2012).

Qualitative approaches

Qualitative methods can bring a range of analytical insight from the nature of the data collected via the social media platforms, and these include (GSR, 2016; Sloan & Quan-Haase, 2016): -

- **Thematic analysis –** Social media data are coded and analysed thematically to identify emotive characteristics of the data or classify the content to find any significant insights within the dataset.

- **Sentiment Analysis (SA) -** A series of specific or tailored algorithms can be used to automatically conduct sentiment analysis to identify whether the text being analysed is positive or negative as regards an entity, such as people, organisation, event, location, or a topic. SA is an active research area, though at present its ability to gauge sentiment for opinions that are ambiguous or complex, sarcastic, inconsistent or contain idioms is somewhat limited.

- **Media analysis:** The audio, video and image content is an important form of online interaction, where relationships can be identified between the content, consumption can be measured, along with the reasons for sharing and reacting/responding to the information.

- **Segmentation/Group identification:** The research community can actively engage with social media data to identify segments that share commonalities and differences with other groups with existing qualitative research. Social media enables researchers to both identify and engage with groups that are hard to reach through traditional methods such as interviews and questionnaires.

- **Active/Passive ethnographic approach:** This approach engages and observes a series of users individually and/or in a group discussion on a social media platform.

The qualitative data analysis may provide an overview of emotions, feelings, and tone, with the influence and power of a topical discussion. This may help identify similarities and differences relating to visual and audio content analysis of photo tags, and the media tone of its content (GSR, 2016; Sloan & Quan-Haase, 2016). Qualitative data analysis can find possible truths behind the numbers providing a higher level of insight (GSR, 2016; Sloan & Quan-Haase, 2016). The main qualitative approach to be used is sentiment analysis. This technique uses a combination of qualitative and quantitative methods because part of the process is manually led to label the training data, so the new data can be accurately classified with a label, assigning it into an emotive category (GSR, 2016; Sloan & Quan-Haase, 2016). The drawback of other qualitative data analysis approaches are that they are based on fewer participants which may not reflect the wider population. The process of coding and theming such data is manually led and is therefore time consuming and requires a lot of manpower (GSR, 2016; Sloan & Quan-Haase, 2016).

The technical tools for acquiring, exploring, transforming and mining the data may consist of using, for example, the SAS, SPSS, and R languages. These data mining tools (notably SAS and SPSS) have an underlying data mining methodology, which are Sample, Explore, Modify, Model, Assess (SEMMA) and Cross-Industry Standard Process for Data Mining (CRISP-DM) (González-Aranda, 2008; Chakraborty, Pagolu, Garla, 2013)

respectively. These two approaches have a common series of stages that are concerned with statistical modelling and data manipulation. These tools are evaluated based on how each is suited to the methodology for the project:

- SPSS Clementine uses Cross-Industry Standard Process for Data Mining (CRISP-DM) which is a comprehensive data mining methodology that was launched in 1996 by SPSS/ ISL and NCR (González-Aranda, 2008; Borges, 2004 & 2011; Wirth, 2000). The CRISP-DM approach considers business objectives, resources, requirements and constraints and has a project management template that is suitable for both large projects and teams to achieve its data mining goals (González-Aranda, 2008; Borges, 2004 & 2011; Wirth, 2000). However, SPSS would be unsuitable, as its "advanced" analytical capability is in fact too simplistic for our purposes.

- SAS has the most powerful analytical capabilities that will meet our requirements to achieve the research aims and objectives. SAS created SEMMA for its data mining software, which can be utilised by other applications, unlike CRISP-DM (Borges, 2004 & 2011; Chakraborty, Pagolu, Garla, 2013; Wirth, 2000). Additionally, our project is on a smaller scale and the large enterprise focused CRISP-DM would not be appropriate (González-Aranda, 2008; Borges, 2004 & 2011; Wirth, 2000). SEMMA can apply various text mining and data mining techniques to gain rich business insights (González-Aranda, 2008; Solarte, 2002).

- R is an open-source language that has no defined methodology (EMC, 2015; O'Neil & Schutt, 2014; Scavetta & Angelov, 2021). SEMMA is more appropriate to apply than CRISP-DM, as it gives R's environment greater flexibility and agility to develop software (EMC, 2015; González-Aranda, 2008; O'Neil & Schutt, 2014). CRISP-DM is restrictive and constraint-led which may have a negative impact on small teams working on packages or applications in R (Borges, 2004 & 2011; O'Neil & Schutt, 2014). To implement a data mining methodology within R may take longer as it does not have one pre-defined, as SPSS and SAS do. R is used most commonly by academic researchers and small to medium enterprises, as it is free (O'Neil & Schutt, 2014). R is growing in the commercialised market, as Microsoft acquired an R-based company called Revolution Analytics in 2015 (DataCamp, 2014; Gartner, 2015). R is a popular language in data science, and with further development in this field large organisations may choose to adopt R (EMC, 2015; O'Neil & Schutt, 2014).

In both the SEMMA and CRISP-DM approaches, there are two ways to gain information from data, comprising unsupervised learning and supervised learning: -

1) Unsupervised Learning (UL) learns from observation and discovery in a dataset (Brown, 2014; Witten et al, 2011; Lin, 2008; Olson, 2008). For example, Unsupervised Learning can automatically find patterns (such as set of class description) and relationships in a given dataset through the application of data mining techniques, such as Cluster Analysis, Rule Association and Kohonen Networks (Brown, 2014; Witten et al, 2011; Lin, 2008; Olson, 2008). Other examples are the organisation of computing clusters, social network analysis, market segmentation, astronomical data analysis and the cocktail party algorithm. The cocktail party problem tries to separate overlapping conversations between each voice/ sound. The analysis can be used to make decisions about the objects that were clustered, or to predict cluster membership for new objects.

2) Supervised Learning (SL) is a form of machine learning, which is a branch of artificial intelligence that concerns the structure and study of systems that can learn from data (Brown, 2014; Witten et al, 2011; Lin, 2008; Olson, 2008). Supervised Learning has known classes and targets sourced from a pre-defined training dataset. A human would provide an algorithm with a dataset to extrapolate "right answers", so for every example in this dataset it has been fed the right option to choose. As a result the algorithm tries to continue giving the "right answers" based on new data (Ng, 2016). This is known as a class description that helps form a classification rule to predict a probability (such as if predicted probability is greater than 0.5 then put observation in class 1) to map new examples of data that has been unseen. SL's common techniques for training data are regression, neural networks and decision trees (Brown, 2014; Witten et al, 2011, 2011; Lin, 2008; Olson, 2008). An example is the development of a diagnostic test, which declares a person to be of class 'healthy' or 'diseased', based on a set of clinical variables (such as blood measurements and medical observations).

3) SL and UL may form semi-supervised learning, which combines SL (labelled training data) and UL (without labelled data) to train (Brown, 2014; Witten et al, 2011; Lin, 2008; Olson, 2008). This may be motivated through supporting predictive modelling at a reduced cost, as labelled data can be costly to generate. This technique typically involves using a small amount of labelled data and a large set of unlabelled data for training, which can improve learning accuracy (Brown, 2014; Witten et al, 2011; Lin, 2008; Olson, 2008).

We will adopt a SL approach as there are known class(es) and target(s), therefore, the UL approach will not be required to observe and recognise patterns to identify its description, class properties and target variables. SL will be developed to predict sentiment on new data. The tools to analyse the data will be discussed next in step 3.4 to ensure the result is accurate and compliant to answer the aim of the project.

3.1.5.1.1 Combining the Approaches

Social media research is “qualitative data on a quantitative scale” (D’Orazio, 2013). Technical aspects with the use of text mining, data mining and sentiment analysis to analyse the data will also be used. This approach may identify significant similarities and differences from both the qualitative and quantitative data to help to establish wider contextual meaning. For example, a selection of Twitter hashtags can lead to a sample being created where the posted language is studied. In this instance, the hashtags can be quantified over time and between different groups, where additional qualitative case studies can help develop an understanding of the hashtag’s use at a particular time.

The development of existing and new machine learning algorithms can perform some human-like actions in an analytical approach with the use of both qualitative and quantitative methods (EMC, 2015; O'Neil & Schutt, 2014; Sloan & Quan-Haase, 2016). These machine learning algorithms are essentially replacing actions that a human would perform, but the outcome is of a qualitative nature. The techniques and processes applied must be viewed separately, such as quantitively on number of users and qualitatively on influence of topics of discussion. The combination of approaches may help to evaluate the analysis of social media data of sentiment analysis in relation to public order.

3.1.5.1.2 Representativity

Social media datasets may be large, as they represent views in real-time and reflect public attitudes that contain links to other online content can allow the researcher to have a wider view on the topic (GSR, 2016; Sloan & Quan-Haase, 2016). Such datasets enable researchers to gain a deeper insight into understanding how conversations are conducted on social media with the use of socio-computational methods (GSR, 2016; Sloan & Quan-Haase, 2016). Social media content can improve understanding of engagement from an anthropological perspective through the analysis of the data to measure public opinion and attitudes (GSR, 2016; Sloan & Quan-Haase, 2016). Research in the social sciences should represent the population of interest, but this is seemingly difficult to determine based on the proportion of Twitter’s active users (Sloan & Quan-Haase, 2016). The large national surveys conducted by the Office for National Statistics (ONS) tend to provide information on the profile and demographics of users (ONS, 2016 & 2021), but it accounts for specific use in time and does not consider the variation of topical conversation.

Twitter does not provide a representative sample of the whole population, so it is important to identify who is represented in the data compared with the offline population (Ruths & Jurgen, 2014). As a result, this will show biases exist and the findings on Twitter may not reflect the same view as other social media platforms and

it may therefore be difficult to infer findings to the general population (GWI, 2015; PEW, 2015; Miller et al., 2015). Social media platform usage can often be dependent on the local context, technological divide and knowledge, cultural, political and social factors (Sloan & Quan-Haase, 2016). For example, Twitter, Google, and Facebook are banned in China (Sloan & Quan-Haase, 2016). Because of these issues, other platforms are most popular in some countries, such as Weibo in China. There are other problems that create problems for reliability and validity, which are outlined below: -

- **Location:** The location of data can be difficult to determine. For example, identifying tweets generated in the UK is a challenge due to the boundaryless Internet (Keith, Ginnis & Miller, 2016). A tweet's geographic location is made harder to identify as only small percentage of users tag their location. If the tweet is not tagged, then it may be possible to identify the user's location using Twitter's meta-data, where 80-90% of tweets are accurately assigned to a location. However, some tweets cannot be assigned a location at all (Keith, Ginnis & Miller, 2016).

- **Spam, Spoofs and Bots:** In 2013, Twitter was estimated to have 10.75 million fake accounts, which accounts for 5% of Twitter's 218 million monthly active users (Yarow, 2013; D'Onfro, 2013). As of June 30th 2017, Twitter had 328 million monthly active users with approximately 82% active users on mobile, and these are the last known published figures (Twitter, 2017b). D'Onfro (2013) suggests 5% of Twitter accounts are fake, which equates to approximately 16.4 million fake user accounts. Fake accounts, such as non-genuine persona and fake bots (tweets not posted by a human) are typically used for automated dissemination and the application of deceptive strategies to influence trending and direct clickstream trails, for example, by the use of a misleading electoral advertisement to influence voters (Cook et al., 2014). Fake accounts can be bought to increase the number of followers to make an individual seem popular, which in-turn encourages other people to follow because of the popularity (Cook et al., 2014). These actions can have a positive or negative impacts on the credibility of social media discourse on Twitter (Cook et al., 2014). Twitter's blue tick feature verifies a user's account, which enables other users to trust that that person's account is credible. In addition, bots can be easily identified due to their posts having a very structured approach (Cook et al., 2014). Social media platforms host automated 'bots', and accounts which pose as genuine human users. GSR (2016) suggests that "large studies should therefore attempt to filter out results from such anomalous sources during analysis." This statement should, however, be taken with caution as this data can be useful depending on the aim of the project.

- **Social media users:** Keith, Ginnis & Miller (2016) suggests social media accounts being analysed can range from powerful users to weaker ones. If

focusing on dominant accounts this might distort the outcome, as it may not capture what everyone thinks on a topic. The balance of individuals and institutions is important when representing public and stakeholder opinions. Twitter does not facilitate an easy way to distinguish between user groups, as certain fields are not mandatory on sign up. This can make it difficult to identify clear distinctions between users (Keith, Ginnis & Miller, 2016). To identify and understand social demography in analysing social media data may help contextualise findings within the data. Socio-demography can help to disaggregate characteristics, such as gender, age, class and occupation, before or after data collection (Keith, Ginnis & Miller, 2016). The demographic profiles of a Twitter user may provide an indication of the reason(s) behind the attitudes expressed in any opinions.

- **Online behaviour:** it can be difficult to understand how reflective a user's behaviour online is of their offline performative actions (GSR, 2016; Sloan & Quan-Haase, 2016). To determine if an online user does align with offline behaviour, additional information is required to identify if there is a differentiation to their profile. Generally, both positive and negative feelings can be over exaggerated online and interest in a topic may not lead to further action (GSR, 2016; Sloan & Quan-Haase, 2016). The personalisation of a user account's preferences can polarise their view, thus limiting exposure to different viewpoints (GSR, 2016; Sloan & Quan-Haase, 2016).

- **Searching for keywords or hashtags:** The relevance and comprehensiveness of the data collected is an important reflection on representativity. Hashtags can be useful for exploring discussions on Twitter, but not all are relevant to a subject topic. Additionally, tweets may go undetected if no hashtag is allocated to tweets relevant to that event (GSR, 2016; Sloan & Quan-Haase, 2016). Hashtags may be avoided deliberately with the wider discussion or expose lack of experience in their use. Hashtags and keywords may not be found in some tweets, therefore the search and extraction of tweets by hashtag may not capture the whole conversation (GSR, 2016; Sloan & Quan-Haase, 2016). The dataset may include non-relevant data to the topic being studied, and relevant data might be missing, which can lead to systematic bias in the dataset. There are other methods of communication, such as non-textual data (images and videos) that will not be captured through data extraction in this project (GSR, 2016; Sloan & Quan-Haase, 2016). We will focus on the use of text, and such non-textual data would require very different analysis techniques.

This demonstrates that bias can be contained within social media populations, as little may be known about the exact population with respect to age, gender, location, level of education and political orientation (GSR, 2016; Sloan & Quan-Haase, 2016).

Therefore, it can be assumed that the data collected may not be entirely representative of the population of the UK. Sampling bias may become an issue as the project limits data in the pre-processing phase, for example, when coding relevant tweets based on the event (GSR, 2016; Sloan & Quan-Haase, 2016). In general, this subgroup will most likely not represent every Twitter user in that dataset. In the next section for 3.1.6, the sentiment analysis approach will be evaluated and an approach will be selected for the project.

3.1.6 Sentiment Analysis Approach

In this section, we will consider sentiment analysis methods in more detail. Figure 3.2 outlines the sentiment analysis methods that will be discussed, including the machine learning approach, lexicon-based approach and hybrid approach. Our eventual aim is to classify tweets as positive, negative and neutral, but to do this in an automated way to adapt to the high volume of social media data.

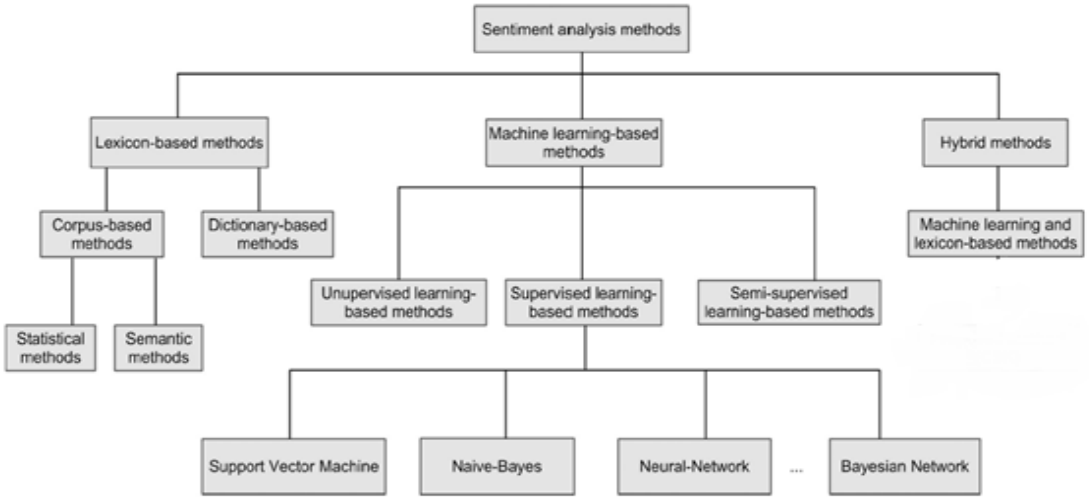

Figure 3.2 Sentiment classification methods

The machine learning technique uses diverse features to construct a classifier to identify the sentiment that a given text expresses (Liu, 2012 & 2015). The use of a supervised learning approach requires a target has to be identified from feature(s), such as polarities as classes that are dependent on target entity or aspect in the sentence (Liu, 2012 & 2015; Medhat, Hassan & Korashy, 2014). The supervised learning approach is one of the most widely used by researchers for its accuracy and adaptability. There are five stages in this approach: data collection, pre-processing, separation into training data and testing data, model creation and validation of results. A model is created on the training set that is then applied to unseen data for classification (Liu, 2012 & 2015; Medhat, Hassan & Korashy, 2014).

The lexicon approach uses different words that are annotated by a polarity score to assess the content to produce a score (Liu, 2012 & 2015). This means that it does not require any training data, but its drawback is that there are some words and

expressions not included in the lexicon which are used within social media posts and elsewhere (Liu, 2012 & 2015). In addition, when dictionaries are created in one specific area (e.g. finance) and applied to other topics, such as politics, then errors can arise, as numerous words have a context-dependent positive or negative connotation (Liu, 2012 & 2015; Medhat, Hassan & Korashy, 2014). For example, "higher crude prices" will be positive in the context of the oil industry but negative in the consumer petrol purchasing context (Liu, 2012 & 2015; Medhat, Hassan & Korashy, 2014). Overall, this approach can show a lack of domain expertise, meaning it is difficult to say that all relevant words and their variants represent a certain concept. The building and maintenance of the lexicon itself can have high impact on accuracy levels (Liu, 2012 & 2015; Medhat, Hassan & Korashy, 2014).

The hybrid approach combines the lexicon-based approach with machine learning techniques to address Sentiment Analysis (Liu, 2012 & Liu, 2015; Medhat, Hassan & Korashy, 2014). This may not be commonly applied, but it tends to show more promising results, as Mudinas, Zhang & Levene (2012) have identified. A lexicon approach first uses the sentiment leixicon to determine the sentiment of sentence or document, and the supervised classifier then uses data that have already been classified by the lexicon as training data (Liu, 2012 & Liu, 2015; Medhat, Hassan & Korashy, 2014). This classifier is then applied to other data to revise the classifications produced by the lexicon. In general, the advantage with this approach is that no data should need to be manually labelled. However, in our case, it was decided to manually classify (label) the data to compare the accuracy of the dictionaries and use dictionaries output of as sentiment category as training data compared with the manual classified (label) category. The main advantage of this approach is that it applies the best of both worlds by using the stability and readability from the well-designed lexicons alongside the high level of accuracy from supervised learning algorithms (Liu, 2012 & Liu, 2015; Medhat, Hassan & Korashy, 2014). Ultimately, the ML approach uses linguistic features, a lexicon relies on sentiment lexicons divided into using statistical or semantic methods to find the sentiment polarity and finally the hybrid approach bridges both approaches (Liu, 2012 & Liu, 2015; Medhat, Hassan & Korashy, 2014). In section 3.1.6.1, we justify the approach used in this project.

3.1.6.1 Justification for Selected Sentiment Analysis Approach

We have chosen to adopt the hybrid approach, combining both lexicon and machine learning approaches to apply sentiment analysis to social media data (Liu, 2012 & Liu, 2015; Medhat, Hassan & Korashy, 2014). The lexicon-based approach will perform at document and sentence level to determine the polarity from the predefined dictionary while the machine learning algorithms (including Support Vector Machine (SVM), Naïve Bayes and Maximum Entropy which will be further discussed in section 5.9) will train a classifier by using the polarity for each sentence as determined by the lexicon (Liu, 2012 & Liu, 2015; Medhat, Hassan & Korashy, 2014). By doing this we can classify

the polarity of other data which can be given the classifier as testing data. We will perform sentiment classification by exploiting training data for each demonstration. This will enable us to identify if a combination of training data performs better than focusing on a single demonstration training dataset (Liu, 2012 & 2015; Medhat, Hassan & Korashy, 2014).

A series of manual, hand classifiers for the sentiment polarity of the sample of each dataset will be carried out to evaluate the accuracy of the sentiment analysis of each lexicon. If it turns out that a particular lexicon performs less well, then this can be removed from further analysis. This combined approach may produce more promising results with respect to precision, recall and F1 measure (Liu, 2012 & 2015; Medhat, Hassan & Korashy, 2014).

We adopt the hybrid approach. As can be seen in Figure 3.3, we use both a dictionary and machine learning approaches in parallel and compare them. On the one hand, we use machine learning to predict sentiment categories with manual classification, and then make a second use of machine learning to make a prediction based on the tweets and manual classification. Finally, in the dictionary approach we compare the gold standard with the majority voting category (e.g., includes all dictionaries).

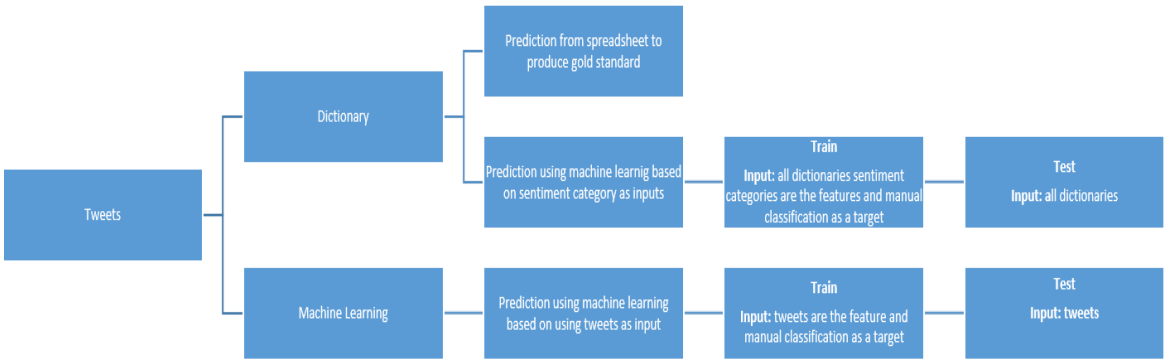

Figure 3.3 Hybrid Approach

In the following chapters, the hybrid approach will be implemented. We describe our process in more detail in section 3.1.6.2.

3.1.6.2 Description of the process

In this section, we will explain the process of the hybrid approach in

Figure 3.4 with a description of each stage emphasised in the following below: -

1. The first step was the extraction of the datasets. DiscoverText was used in this instance to extract the data from Twitter's API with the use of set keywords as outlined in section 3.1.3.

2. The data were then reviewed through exploration in order to understand the datasets. The datasets were initially manually coded (relevant) data, where a keywords list was created for each dataset to filter out the relevant tweets to speed up the process, but this may remove a limited number of tweets that are relevant. Each of these lists was extended with words from their respective dataset. These additional terms are identified with the use of Term Frequency (TF) and Term Frequency Inverse Document Frequency (TF-IDF) that are common document clustering methods that that try to reflect how important word is in a document in a corpus, especially with short texts (Aggarwal & Reddy, 2014). Each extended word list was fed to the automated process to filter out the most relevant tweets from the remaining tweets from each dataset.

3. Pre-processing then involved applying a series of techniques to the data to reduce the noise within the text and lessen dimensionality to improve classification effectiveness. A series of papers (Ghag & Shah, 2015; Haddi, Liu & Shi, 2013) have used a standard stop word list to remove common words that have no bearing on the semantics of a text and extended them further when pre-processing their datasets to help improve the performance of a model. However, there are researchers such as Saif et al (2014) who suggest that a dynamic generation stop word list is far more effective standard stop words can have a negative impact on the sentiment score, the levels of which can vary dependent on the dataset. A standard stop word list in R was adopted when pre-processing the datasets, as on examining the data, it appeared that using this list would not make such a negative impact. This decision will be evaluated to determine if this was the most effective approach to sentiment analysis.

4. **Classification / Dictionary Based Approach**
    a. **Dictionary Based Approach:** The manually coded (relevant) data and automated coded (relevant) data will both be processed through the sentiment dictionaries (Liu, 2012 & Liu, 2015; Medhat, Hassan & Korashy, 2014). In the feature selection, the most relevant attributes will be selected, and the features will be extracted, with combined attributes formed into a newly reduced set of features. The polarity detection will be applied at the sentence level to determine each individual tweet's overall sentiment (Liu, 2012 & Liu, 2015; Medhat, Hassan & Korashy, 2014).

    b. **Classification:** The bank of keywords created in the coding phase decided to use the bank of words to apply this as an automated process to seek out relevance in the whole dataset, which can be used to classify the entire dataset (Liu, 2012 & Liu, 2015; Medhat, Hassan & Korashy, 2014). We have a subset and the entire dataset. The subset manually coded will be used as train/test/validation set for the machine

learning process, then the resulting model will be applied to the entire relevant dataset to predict the sentiment (Liu, 2012 & Liu, 2015; Medhat, Hassan & Korashy, 2014).

c. Both a and b above both require a form of evaluation as specified in section 3.1.8. As a classification problem, Sentiment Analysis uses the evaluation metrics of Precision, Recall and F-score (refer to section 5.10) (Liu, 2012 & Liu, 2015; Medhat, Hassan & Korashy, 2014). In addition, the lexicon approach uses measures, such as macro and micro averages (Liu, 2012 & Liu, 2015; Medhat, Hassan & Korashy, 2014).

5. The results of both dictionary approach (input dictionaries and manual classification) and highest performing algorithms in the tweets and manual classification machine learning approach will be examined in the change point analysis.

6. Change point analysis will be applied to attempt to identify any significant change over time in sentiment.

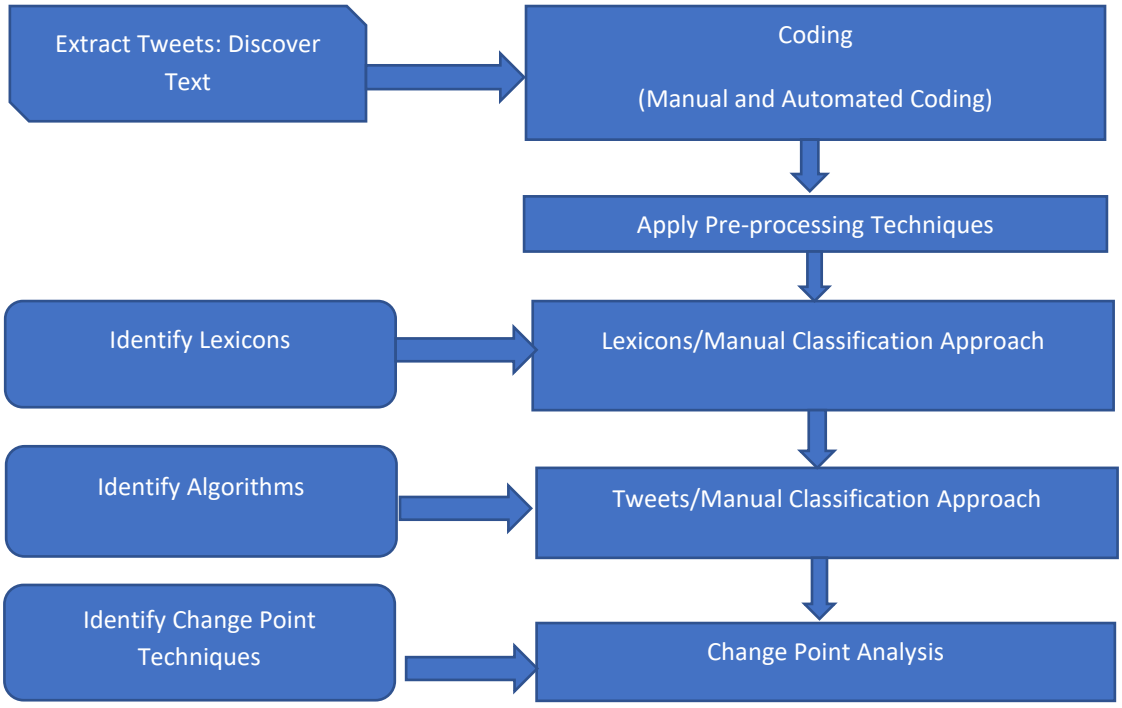

Figure 3.4 Hybrid approach's process

In this step of the analysis phase there has been consideration of ways to analyse social media data and which sentiment analysis approach will be adopted for the project. The tools and techniques established in each of these steps will be applied to the methodological approach that will be followed in the implementation phase, which is discussed in step 6.

3.1.7 Step 6: Implementation

The project will help towards building a model to predict what may happen in an event to prevent disorder and increase public safety. This may help notify communications teams in the Law Enforcement Agencies, who will use the early warning to decide whether to enact an intervention to reduce to prevent public disorder.

The implementation of the proposed framework, some of which has been discussed above already, will be applied in section 4 to 6. We will detail the evaluation techniques used on the social media data analysis in section 5.10 which include precision, recall and f-measure to determine the reliability of the results. In step 7, the project framework with the rest of the project will be evaluated.

3.1.8 Step 7: Evaluation

The project will be evaluated from how it relates to aim, objectives, deliverable, and the framework, which will be outlined in section 7. After the evaluation, then it's about managing the knowledge, which is detailed in step 8.

3.1.9 Step 8: Knowledge Management

The work conducted in this area (as outlined in first part of this chapter) shares knowledge on social media research methodology based on the pilot study in form of a publication (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018). Additionally, a multidisciplinary training session on social media analysis happened based in the doctoral school. Further work will be outlined once the project has completed the social media research life cycle based on the demonstration case studies.

The social media strategy established in this chapter is the framework to be followed in the implementation phase in sections 4 to 6.

4 Pilot Study and Lessons Learned

To prevent potential problems with the proposed project, a smaller scale, pilot study was conducted beforehand. This pilot study would expose potential, unforeseen issues prior to the start of demonstration case studies, meaning appropriate solutions can be put in place to improve the sentiment analysis outcome. The study will provide a set of recommendations on the most optimal tools and techniques to extract, transform, analyse, and visualise in the sentiment analysis process. These recommendations will help to enhance the demonstration case studies results in section 6.

4.1 Baltimore Riots

The pilot study focuses on the Baltimore riots as the researcher collected data at the time of the event between April to May 2015. Since then, the Baltimore riots as an area of research is widely known (Choudhury et al., 2016; Fichet et al., 2016; Korolov et al., 2016; Marshall & Wang, 2016; Wang, Marshall, Huang, 2016; Zou & Song, 2016). Some of these papers have used either the exact same hashtags/ keywords to extract the data and it appears that these authors have bought the data retrospectively, whereas the small sample of data for the pilot study has been extracted live from Twitter.

An application called 'hashtagifme' was used to identify the relevant hashtags. As a result, the most relevant hashtags are "#FreddieGray" and "#BaltimoreRiots" and "#BaltimoreProtests" based on the Baltimore event. These hashtags were applied in the NVivo software to capture data. This software is too restrictive in extracting all its value from the data, so it was exported as an Excel file to import into R and Tableau. In the early stages of the project it was difficult to determine which text and data mining techniques are appropriate due to lack of experience. Despite this, different techniques are explored at length to determine relevant ones to help answer the research question.

In the development phase, relevant R packages are researched and identified to load in the data and cleanse it as specified in section 5.8, which are 'TM' (Text Mining), 'NLP' (Natural Language Processing), 'stringr' (remove characters), GGPLOT2 and 'wordcloud' to assist in visualising the data. To aid in the development, code examples of social media data mining were examined online via 'RPubs', 'RBloggers' and 'Towards Data Science' communities and in books (Lantz, 2015; Kwartler, 2017; Silge & Robinson, 2017) to determine which functions are best utilised for this process. Using this acquired knowledge, a corpus is created, and features are extracted from pre-processing text for word frequencies and complex analytical tasks, such as sentiment analysis.

The collected data is visualised to gain a deeper understanding of the domain and identify any abnormalities before any transformation. Figure 4.1 displays many tweets by hour with peaks and troughs over time. Additionally, there is a problem with Figure 4.1, as the timeline is consistent with major declines in its trajectory which is due to the different times when the data is collected. These dramatic drops from each peak, from the 1st of May onwards show less tweets because this takes place after the event. Furthermore, these tweets are not consistently collected on the day. Therefore, it makes this timeline unreliable. This emphasises the importance of having a constant stream of data or buying the dataset directly from Twitter to have a wider understanding of the situation.

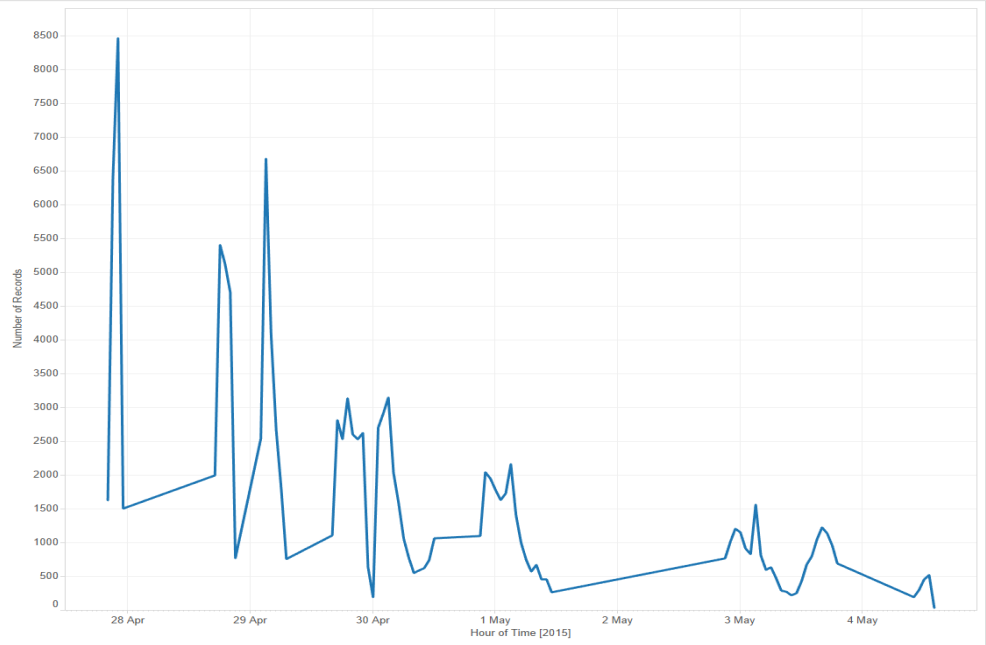

Figure 4.1 Number of tweets over time by hour

Figure 4.2 shows a series of bar charts displaying 'Total tweets by username', 'Total tweets by username for event' and 'Total retweets by username for event'. The 'Total tweets by username' is the number of tweets since the user's account was registered. This helped to identify users who overall had most tweets since opening their accounts. This profile data was of less importance as we are focused on the discussion within the event. However, the 'Total tweets by username' shows the top 10 users are the most prominent tweeters than the remaining users, but the retweeted category top 10 are a different set of users except for 'PulpNews' and 'I_Cant_Breathe_'. Additionally, 'I_Cant_Breathe_' posted the most tweets for the event, but none of the top usernames appear in high up in the number of retweets for the event.

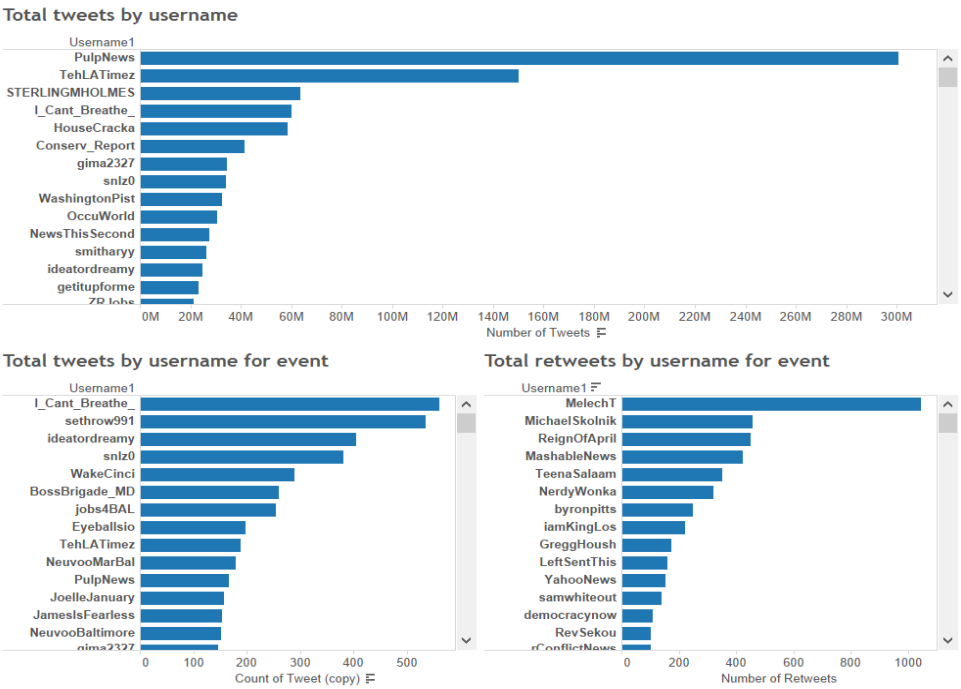

Figure 4.2 Dashboard tweets and retweets by username

The next exploration will involve identifying the prominent keywords based on the Baltimore riots. As predicted, 'baltimore' is highest on the list of keywords and its dominance skewed the scale of the bar chart, so this was removed to help rescale the bar chart's proportional representation in Figure 4.3 to provide a fairer representation. The other words are lower by at least 50,000 compared with 'baltimore'. This keyword along with others are highest due to the biased way of collection of the data with the hashtags used to extract the data. These hashtags are the main keywords used in most tweets due to the nature of the event. Therefore, the result was always going to be skewed.

The words 'baltimore' (mentioned 134,316 times), 'freddiegray' (14,467), 'baltimoreriots' (12,355) and 'baltimoreuprising' (9,566) are excluded from the bar chart as they are unique to the event. Additionally, stop words will not be removed as there is a concern important words could be removed changing the context of the outcome. Some of the highest top terms identified are 'police', 'blacklivesmatter', 'riots', 'curfew' and 'black'. These words commonly appear, thus being the main topic of discussion online, which is further evidenced by observing the live stream of the event on periscope. The use of prominent words may help identify topics through the timeline of a live event. This may enable the police to understand when to interject to maintain the peace. Other techniques will be explored beyond the pilot, such as a log scale to represent all terms to their exact scale rather than removing words.

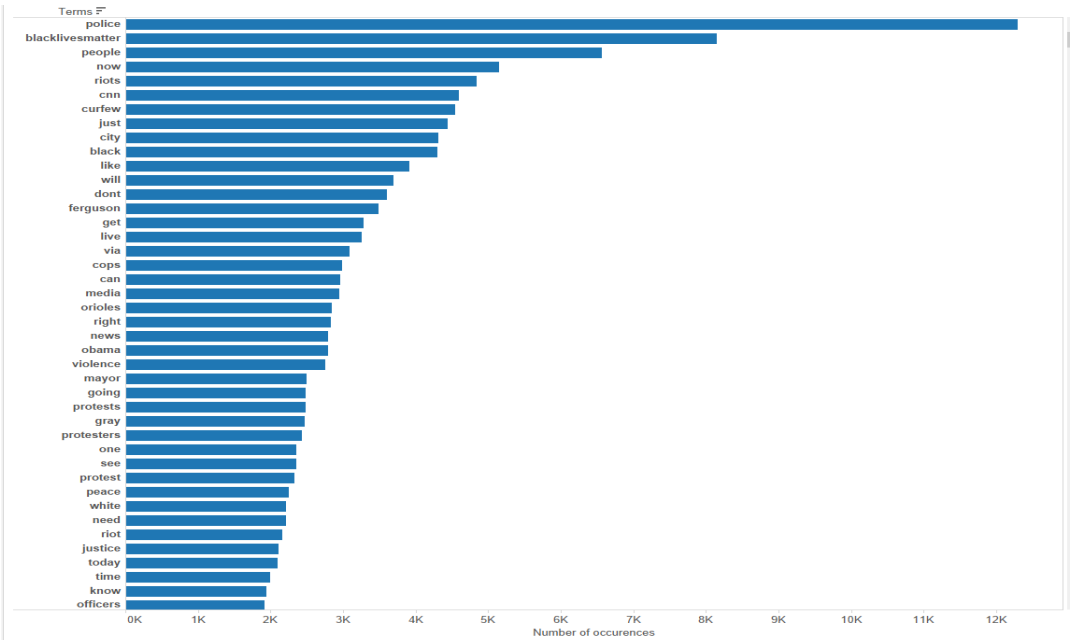

Figure 4.3 Terms by number of occurrences from all tweets

Figure 4.3 helped to identify the most popular terms, but it can be difficult to visualise the event. Chen, Lin & Yuan (2017); Cho, Wesslen & Volkova (2017); Kavanaugh et al (2012); Nazer et al, (2017); Ragini, Rubesh Anand & Bhaskar (2018) have used word clouds as an effective way to describe an event, which is applied to this pilot. Figure 4.4 has applied the cleansing techniques without the removal of stop words, which

shows less relevant words to describe the event. In Figure 4.45, the cleansing methods with TM package's 174 stop words (Feinerer, Hornik & Artifex Software Inc, 2018) with additional unique and irrelevant words removed has provided a wider view of the event. The word cloud might be a useful indicator to identify what is happening at an event live at the time, instead of analysing the data without a filter on the stop words.

Both Figure 4.4 and Figure 4.5 identified other keywords that will help code the data based on its relevance to the event, which may speed up the process of coding. Other techniques will be explored to improve the accuracy of the result, such as using Term Frequency-Inverse Document Frequency (TF-IDF) to reflect on the importance of the word to a document in the corpus and possibly use of Zipf's law for automatic generation of stop words (Lo, He, Ounis, 2005; Saif et al., 2014).

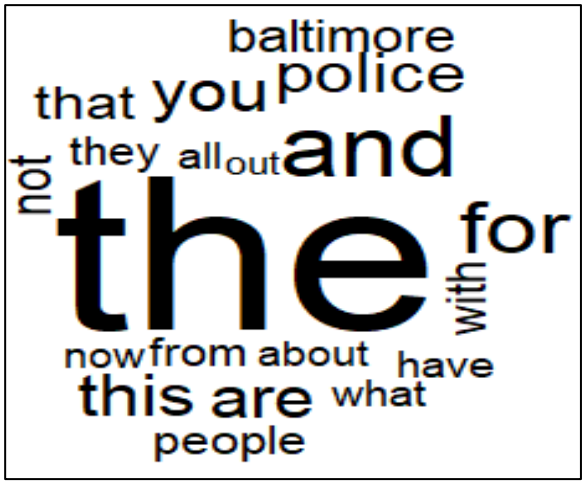

Figure 4.4 word cloud with no filter

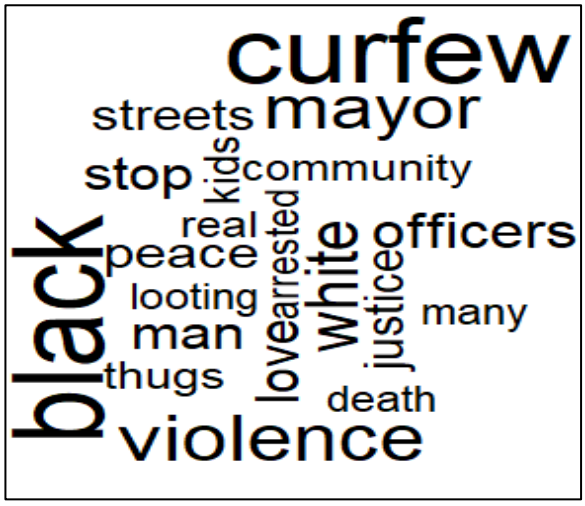

Figure 4.5 word cloud with a filter

[Intentionally Left Blank]

In the next step, the untransformed data is used with the “sentiment” package created by Timothy Jurka (Jurka, 2012). When the pilot study concluded this package was not available on CRAN. R was used within Tableau to detect sentiment and visualise the results, but this presented a series of drawbacks, such as R took longer to produce the outcome for Tableau to output the visualisation. As alternative approach, R is used to detect the sentiment within RStudio, and then exported to Tableau to speed up the process. Figure 4.6 shows total tweets by sentiment category by day. These results have a higher positive rate than the other two sentiment categories.

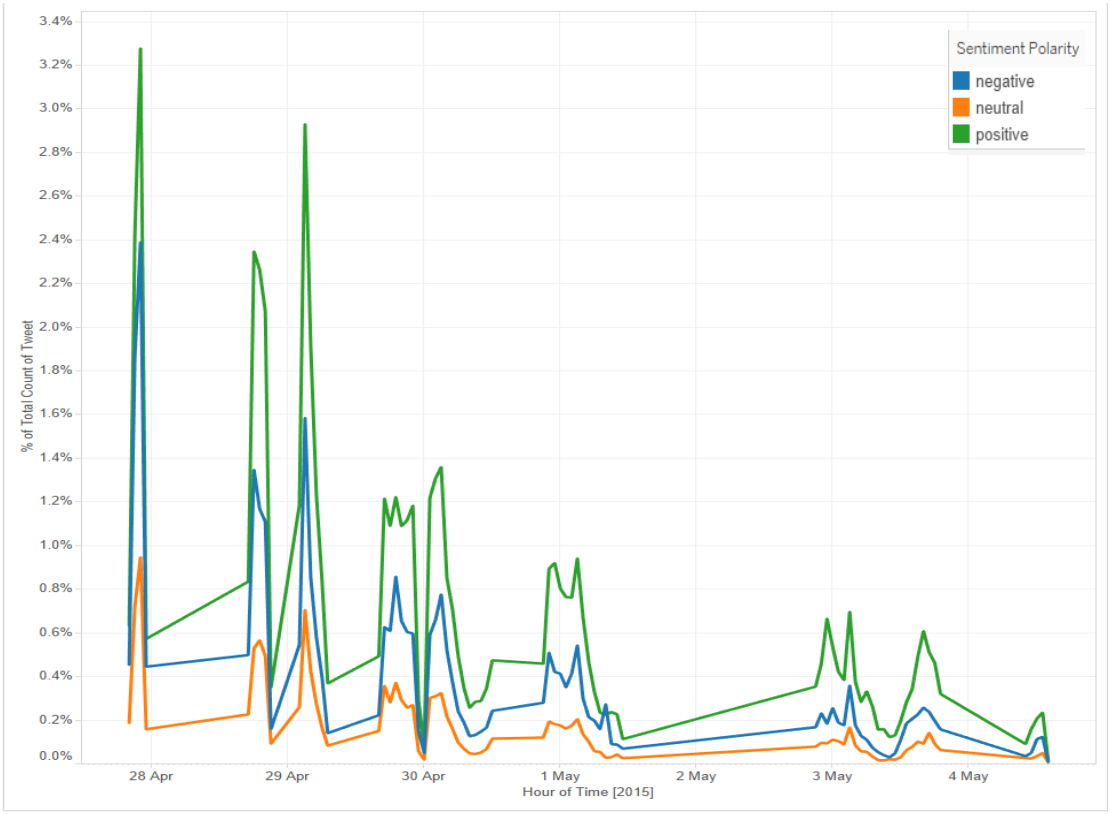

Figure 4.6 Total tweets categorised by sentiment

To validate results of the sentiment, Figure 4.7 shows a dashboard that contains a timeline of the event by sentiment with a list of tweets. The highest peak is selected, which filters relevant tweets in the list for that period of time. The tweets listed in Figure 4.7 classified as “positive” by the dictionary on observation were found to belong to a different category, such as “neutral” or “negative”, which highlights some tweets were misclassified. For example, the tweets containing “@SkyNews..” are neutral and the one with “#Baltimore Police: Gangs have entered into a partnership to “take out” law enforcement officers….” is negative as it refers to gangs harming police officers. As a result, there is a need to compare this sentiment outcome with other dictionaries to validate the accuracy of the result. Additionally, there is a large number of studies use an evaluation technique called precision, recall and f-measure (refer to section 5.10) to verify the accuracy. This method will be employed with the demonstration case studies. Furthermore, manual classification will be conducted for the case studies to check the algorithms’ reliability and this labelled data will be used

for machine learning to train and test data to validate the results. Moreover, machine learning will enable the labelling of the rest of the tweets that may reach an insight or recommendation.

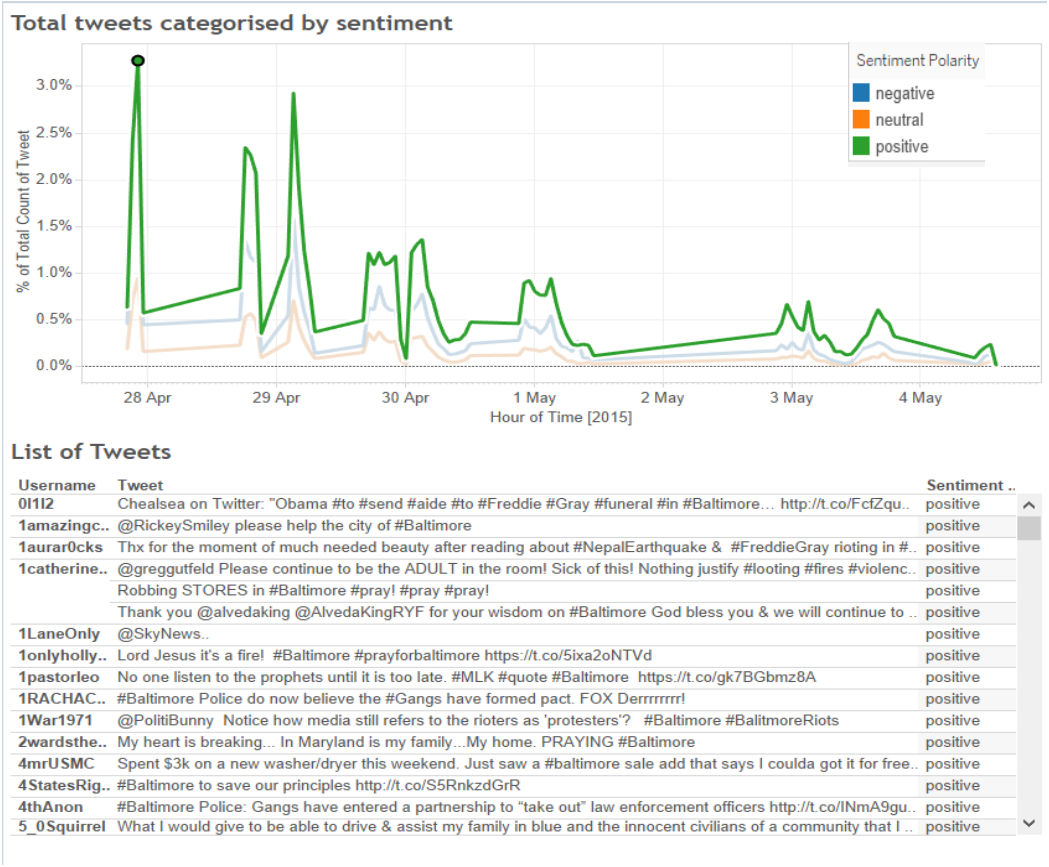

Figure 4.7 Filtered list of tweets by total tweets categorised by sentiment

The "sentiment" package included another function that can breakdown sentiment by emotion rather than polarity as depicted in Figure 4.8. This shows "Null" has the majority over the other emotions listed, such as anger and joy.

Figure 4.8 Proportion by emotion and related tweets

There are fine margins to what is anger, joy or fear. Therefore, the result is more likely to be “Null”, which could be due to the tweets containing no emotion, as shown in polarity as being “neutral”. The dictionary’s limitations may not contain enough words that are within these tweets. Therefore, these tweets are unlikely to be placed into a category of emotion. Bermingham (2009), Cohen et al (2014) and Jurek, Mulvenna, Bi (2015) show most studies use of dictionaries focus on polarity rather than breaking it down into several different emotions, as it can be difficult to classify the results with less accuracy. The focus of the case studies will be on polarity.

In this pilot study, change point analysis has been researched from a theoretical perspective and technical level to find relevant R packages for change point (refer section 4.1). The initial R packages identified are “changepoint”, “BCP” Bayesian Analysis of Change Point Problems, “ECP” Non-Parametric Multiple Change-Point Analysis of Multivariate Data, and “CPM” Sequential and Batch Change Detection Using Parametric and Nonparametric Methods which may be used in the analysis. Some techniques identified, such as Latent Dirichlet Allocation (LDA) and Named Entity Extraction in the pilot study are no longer used as the project’s focus is not on situational awareness. The pilot study identified positives in our approach and highlighted a series of problems with a series of proposed solutions when applied to the case studies.

In section 5, the UK demonstrations will be explored and analysed providing background information with the techniques learnt from the pilot study.

5 Initial Data and Information Processing

The focus of this chapter is to explore the case studies data to gain a greater understanding of the data and to determine the best approach for pre-processing tweets to prepare for application of data mining techniques to then provide insights into the data and hence our research aims.

5.1 Metadata Composition

The four datasets acquired from Sifter each contain 276 columns. One tweet provides a large number of fields, some of which are of interest e.g. text of message, date, time and entities that will be consistently analysed (Sloan & Quan-Haase, 2016). An entity is metadata and additional contextual information within a tweet, which contains hashtags, URLs, media fields and user mentions (Sloan & Quan-Haase, 2016). The datasets are downloaded in 50,000 tweets per block from Sifter, in line with Twitter’s service agreement. This might make it more manageable for some applications (e.g. Microsoft Excel) to process the data. The four datasets are comprised of 565,000 tweets in total.

The research is focused on the actual tweets, but other data associated to the posts is important. In the following sections, the tweets and fields are explored on language, location, date/ time, retweets, bad data and coding relevant tweets with use of keywords. As a result of this exploration, some tweets and fields related to the tweets may not be required this project.

5.2 Language

Sloan et al (2016) have suggested that 40% of Twitter content is produced in the English language. The researchers sampled 113 million tweets, 33% of which were different to the user's selected language. The "language" column in MMM 2015 dataset outlines that "en" (English language) 151,760, but "en" by actor language is 144,747 (other popular languages in line are Spanish, French and German), both out of a total of 181,711 tweets. MMM2016 shows that "language" has "en" of 93,169 and "actor_languages" is 90,401 of "en", which both are out of 108,456 (other popular languages in line are Spanish, German and French). Dover 2016 shows that "language" called "en" is 17,923 and "actor_languages" is 15,515 of "en", which both are out of 25,031 (other popular languages in line are German, Italian, Greek, Spanish and French). Anti-Austerity 2016 shows that "language" called "en" is 219,977 and "actor_languages" is 207,543 of "en", which both are out of 250,416 (other popular languages in line are Spanish, French and German). This shows most tweets are in English, but this is largely due to the localised nature of each topic for the event.

The analysis of language is complex on Twitter as members can be of different nationalities communicating in different languages. Even when filtering based on 'en' in the fields actor_languages and language (user's selected language) fields, the results can contain tweets with a different language. When a user registers, Twitter provides an option to select a language of preference, but actor_languages refers to the language of the tweet (Sloan & Quan-Haase, 2016). Furthermore, language can be determined by the type of users mentioned, language specified via the hyperlinks, hashtags and the language of the original post. A user may be multi-lingual, and therefore, the language used when posting tweets may not match their registered preference language. These different options can make it difficult to identify and categorise based on the user's primary language. The decision has been made to lessen this problem by removing non-English tweets. Had this been a problem, then we could have attempted to build in capacity for multilingualism (Sloan & Quan-Haase, 2016), but this did not prove necessary. Initially the possibility of using applications to translate the language was considered, but we decided against the approach as the inaccuracies of translation can change the context of a situation substantially if incorrectly translated. Overall, using multiple languages in the research was overly and unnecessarily complicated, due to resource limitations this type of research cannot be conducted. This is due other language specialists being required to address this methodological issue to ensure accurate results. Sentiment analysis techniques will

therefore be applied to detect the emotion expressed in the English tweets. The English words in the tweets could be written in UK English or other forms of English language, such as American English. This may cause an issue for the UK/US dictionaries to identify words in the tweet to classify its sentiment as positive, negative or neutral.

Further, although in the field of sentiment analysis, development of dictionaries is mainly centred on American English language more than any other language (Sloan & Quan-Haase, 2016). As a result, dictionaries available in other languages and even other varieties of English are limited, with a need to build on existing languages or to create a dictionary for different language to score data. A creation of a dictionary requires specialist linguistic skills. There are multiple complications involving other languages than English, hence removal of all non-English tweets.

5.3 Location

The geo-coordinates of tweets do not appear often in a posted tweet, hence cannot be relied on for mapping tweets. In addition, this still can be manipulated if a user understands how to technically change the location of their phone from, say, the UK to Switzerland when posting a tweet. The '[M] user_location:' field could indicate possible locations of tweets for many of the tweets, but still there are a high number of blanks for this category. However, user location can be falsified as well and can be misleading as the tweet may not be coming from the user-registered location. Moreover, this location data can be inconsistent. for instance, individuals can code their location differently, which is evidenced from specifying London as "London, UK" and "London, England" in the 2016 Anti-Austerity dataset.

5.4 Date and Time

Time is an important dimension to analyse because tweets are temporal and tend to have to a short time span for relevance, though this can vary depending on the longevity of a topic (Sloan & Quan-Haase, 2016). Demonstrations may be built up days before the event has arrived to draw interest into the topic. After the event has occurred, the topic lessens in interest (Sloan & Quan-Haase, 2016). There are some exceptions where topics are consistently debated, such as #BlackLivesMatters. These topics may increase in tweets once a similar incident arises again that brings this topic back to the forefront in the public sphere (Sloan & Quan-Haase, 2016). Jungherr & Jurgens (2013) tried to identify trends to understand the dynamic structure of broadcasting events and the persistence of spikes when specific terms are used in an event. The "[M] posted_time:" field is the origin for recording of time and date of each tweet. Other numerous fields of time have been removed as many have no record of time or are repeating the posted time within another field.

5.5 Retweets

A retweet is a repeat of the same original tweet. This is either produced by a user or bot. In the dataset there is a field called “Is_retweet”, which is propagated by the user selecting the “retweet” button. This is not completed by all users and some follow a convention of using “RT” or “VIA” or an alternative way when duplicating the original tweet. The removal of these tweets can be speeded up by alphabetically sorting the tweets, as this makes it quicker to identify and delete the retweets.

Boom, Canneyt & Bart (2015) has shown that some researchers remove duplicates from their overall dataset, dependant on what they are trying to achieve such as reducing the amount of data for code for theming. This enables a researcher to save time to focus on other areas of interest, which they may not be able to do without reducing the data (Boom, Canneyt, Bart, 2015). These duplicates are going to be retweets of the original post, which is of less value. The number of retweets can be calculated to identify how much influence that original tweet has within the community, then the retweets will be removed.

5.6 Bad Data

There is no clear definition of what bad data is, but it can be considered a technical phenomenon that includes: missing values, malformed records, incorrect values and inaccurate or irrelevant parts of data (McCallum, 2013). Missing data occurs when a piece of information exists, but has not been included for in the raw data for some reason. An example of this could be missing words in a sentence that could change its meaning or a numeric value being blank or a missing value being substituted by the value of zero. Data collected from Twitter users may contain users who do not want to provide information, leading to missing values. For example, geographic co-ordinates may not be completed due to a user’s privacy settings and their profile bio might not be fully completed. In addition, data may be incomplete due to the application, for example, automated tools (e.g. TAGs) using Google spreadsheets have document limits on the amount of data that can be collected. The data captured in a time block can vary with Twitter API restrictions and exclusion of deleted tweet(s). The level of incompleteness in the dataset can vary, thus a few lines of enquiry may not be able to be explored, which can directly impact the result of analysis.

Incorrect data occurs when part of the information has been incorrectly specified, for example, an error on choice of word or placement of a decimal point or being incorrectly interpreted, such as assumptions about whether the text is US, rather than UK, English. In addition, data inaccuracies can occur when users input data that may be untruthful. Furthermore, there could be inconsistencies with formatting of data, such as different date formatting, and paragraphs of text containing a mixture of two

languages rather than being in one consistent language. The date format is consistent between tweets as Twitter uses a default format for any tweet posted worldwide.

There are many examples of bad data in our datasets and these will be cleansed in the pre-processing phase to prepare the data for the analysis stage.

5.7 Coding with keywords

Relevant keywords associated with each demonstration that are unique or common are identified. Some words can be themed into topics to give the list order, but also make it easier to identify other words that can be searched to keep relevant tweets or remove irrelevant data. A list of words will be identified for each dataset to code tweets into either irrelevant or relevant categories in relation to the demonstration in question.

In section 5.8, the relevant tweets will then be cleansed to prepare for the analysis phase.

5.8 Data Cleansing

In this phase, the most important fields will be selected out of the 276 from the Twitter datasets. Based on the knowledge acquired in this project we determined that 49 of the 276 are the most important. The remaining 227 tend to be mostly incomplete or replication of other data presented in different fields, thus making them less significant or simply duplicated information. In addition, these fields omitted are less relevant to the project, as the focus is on the actual tweets.

The dimension of a feature vector can be large even for relatively small documents such as tweets. Some elements can be dropped without affecting the performance of the sentiment analysis outcome. Textual data often provides inconsistencies that might cause algorithms to glean inaccurate insights from the data. Feature selection is a process to remove irrelevant features and any irregularities. Along with this, it also reduces the size of the vector and this reduces computational time, which may lead to improving the performance of the analysis resulting in a higher level of accuracy in the results (Silge & Robinson, 2017). The text mining process is to clean the data for preparation towards the data mining phase to understand the patterns and trends from the data (Silge & Robinson, 2017).

In the text mining phase, R packages will need to be identified (Silge & Robinson, 2017) to load and pre-process raw data in R (the importance of text mining is outlined in section 2.4 of the literature review). The text mining techniques are chosen on the

basis of research into the common approach of how Twitter is text mined on RPubs and in a series of research publications/books (as emphasised in both sections 2.4 and 2.5 the literature review), so here are the techniques as follows: -

- Remove "amp;", and "\n" which are not part of the tweet
- Remove html links, which are not required for sentiment analysis
- Remove retweet entities from tweet
- Create a list of hashtags and emoticons for record
- Remove all "#Hashtag", "@people", punctuation, emoticons
- Remove numbers, we need only text for analytics
- Remove unnecessary spaces (white spaces and tabs)
- Remove extra characters
- Remove stop words and unique words for each event
- Convert text to lower case

The reduction of the feature set may bring limitations to what is presented in the analysis stage. It should be noted that inclusion or omission of these techniques may influence any outcomes (Silge & Robinson, 2017). Therefore, it is important to select words to remove that will not have a major impact on the results' accuracy, as it may change the outcome of a tweet having a positive, negative or neutral sentiment (Silge & Robinson, 2017). The data will then be explored to further understand the domain and to provide background of each event, such as total tweets over time and frequencies of the most important tweeted words.

Once this general background information is gathered, the data mining phase will begin.

5.9 Data Mining Approach

In the data mining approach, the sentiment analysis technique is applied to detect emotion from the textual data to understand whether the demonstrations may reach a higher level of tension and potentially turn a peaceful demonstration to a potential riot. A supervised approach is being adopted as the target is already identified as outlined in section 3.1.5.1.1. A range of lexicon dictionaries are available in the English language sourced from within R packages except for SentiStrength which is standalone, and these are outlined in Table 3.

Resource	Entry size	Sentiment category and score range	Notes
Sentiment dictionaries contained in Lexicon package			
Jockers	10,738 words	Sentiment values ranging between -1 and 1.	Dataset containing a modified version of Jocker's (2017) sentiment lookup table used in Syuzhet.

Jockers Rinker	11,709 words	Sentiment values ranging between -1 and 1.	Dataset containing a combined and augmented version of Jockers (2017) & Rinker's augmented Hu & Liu (2004) positive/negative word list as sentiment lookup values.
Huliu	6874 words	Sentiment values (+1, 0, -1.05, -1, -2)	Augmented version of Hu & Liu's (2004) positive/negative word list as sentiment lookup values
SentiWordNet	20,094 words	Sentiment values ranging between -1 and 1.	SentiWordNet ver. 3.0. Based on WordNet 3.0 (Baccianella, Esuli, Sebastiani, 2010; Esuli, & Sebastiani, 2006).
NRC	5468 words	Sentiment values of either +1 and -1.	A filtered version of Mohammad & Turney's (2010) positive/negative word list as sentiment lookup values.
Loughran Mcdonald	2702 words	Sentiment values of either +1 and -1.	Financial word list as sentiment lookup values (Loughran & Mcdonald, 2016)
Senticnet	23,627 words	Sentiment values ranging between -1 and 1.	Augmented version of Cambria, Poria, Bajpai,& Schuller's (2016) word list as sentiment lookup values.
Inquirer	3450 words	Sentiment values of either +1 and -1.	Based on Harvard IV-4 and Lasswell Dictionaries (Harvard, 2002).
Slangsd	48,277 words	Sentiment values ranging between -1 and 1.	Dataset contains filtered version of Wu, Morstatter, & Liu's (2016) positive/negative slang word list as sentiment lookup values. All words containing other than "[a-z ']" have been removed as well as any neutral words.
Socal Google	3290 words	Sentiment values ranging between -31 and +31.	Version of Taboada, Brooke, Tofiloski, Voll, & Stede's (2011) positive/negative word list as sentiment lookup values.
Vadar	7236 words	Sentiment values ranging between -1 and 1.	Dataset containing a filtered version of Hutto & Gilbert's (2014) positive/negative word list as sentiment lookup values.
Sentiment dictionaries contain in Syuzhet package			
Syuzhet (default)	10,748 words	Sentiment values ranging between -1 and 1.	"Syuzhet" lexicon is developed in the Nebraska Literary Lab under direction of Matthew Jockers (Jockers, 2017). This lexicon created from 165,000 human coded terms from corpus of contemporary novels.
Bing	6789	Sentiment values (+1, -1)	The "bing" lexicon was develoepd by Minqing Hu and Bing Liu as the OPINION LEXICON See: http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
AFINN	2,477 words	Ranging between - 5 (very negative) and 5 (very positive).	Based on Affective Norms for English Words (Nielsen, 2011).
NRC (NRC Word-Emotion	14,182 words	sentiments: negative, positive	Based on Mohammad & Turney (2010) paper called "Emotions Evoked by Common

Association Lexicon)		emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, trust	Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon."
Independent packages			
Sentiment Berkeley (R package deprecated)	1542 6518	Positive/negative/neutral also anger, surprise, joy, etc.	R package called "sentiment" has Bayesian classifiers for positivity/negativity and emotion classification (Jurka, 2012)
Stansent	Unknown	Sentiment values ranging between -1 and 1.	This dictionary is a re-implementation of Matthew Jocker's Stanford coreNLP wrapper in syuzhet (Jockers, 2017; Rinker, 2017). The R pakcage stansent wraps Stanford's coreNLP sentiment tagger. Tag sentiment as most negative (-1) to most positive (+1) (Rinker, 2017).
SentiStrength (not an R package, separate application)	2546	Ranging between - 5 (very negative) and 5 (very positive).	SentiStrength is a tool that is constructed by combining GI and LIWC dictionaries and includes lists of negations, intensifiers and emoticons (Islam & Zibran, 2017; Thelwall, 2019).

Table 3 List of Sentiment Packages

In total, 18 lexicon-based dictionaries will classify the relevant cleansed tweets. Additionally, there will be a 19th dictionary which combines several dictionaries to identify whether a larger dictionary can improve how the classifier determines the outcome of positive, negative and neutral scores. These dictionaries exhibit different ranges of positivity and negativity, with scales ranging from -1 to +1 and -5 to +5. Some dictionaries, such as Hu Liu and Bing Liu range is different, but the scales indicate a similar output, such as -0.26 instead is -1 or 0.5 is 1. Therefore, the difference in the outcome is not significant when the scores a rescaled in the same range. The majority of these dictionaries are American English, with the exception of one called "SentiStrength", which is UK English. A combined dictionary will be formed that is made up of 11 lexicon-based dictionaries. These dictionaries are chosen on the basis that if the scored word list is similar then these lists could be obtained to combine the dictionaries. The combined dictionary will have its sentiment scores standardised within a specific range of -1 to +1, then the words in the dictionaries can form into one large sentiment score list. This combined dictionary will be compared to the other individual 18 lexicon dictionaries.

A series of ML algorithms will be deployed to infer the sentiment label, which will be informed by the unsupervised approach's sentiment scores for manually annotated data (Aggarwal, 2015; Aggarwal & Reddy, 2014; Lantz, 2015). There are supervised algorithms used for ML that can be applied to the datasets, which are: -

- **Tree:** A decision tree is a graph has a series of branches to illustrate every possible outcome of a decision (Aggarwal, 2015; Aggarwal & Reddy, 2014; Lantz, 2015). This is a way to simplify a complex strategic challenge(s) and to evaluate the cost effectiveness of a decision.

- **Random Forest:** This combines individual decision trees together, as it can strengthen the predictions outcome compared with using a single tree (Aggarwal, 2015; Aggarwal & Reddy, 2014; Lantz, 2015).

- **Naïve Bayes:** This is a probabilistic classifier with a strong conditional independence assumption that is optimal for classifying classes with highly dependent features. Adherence to the sentiment classes is calculated using Bayes' theorem (Aggarwal, 2015; Aggarwal & Reddy, 2014; Lantz, 2015).

- **Max Entropy:** This is a probabilistic classifier belonging to a class of exponential models (Aggarwal, 2015; Aggarwal & Reddy, 2014). Unlike the Naïve Bayes classifier, Max Entropy does not assume that its features are conditionally independent of each other. Instead, Max Entropy is based principally on the Maximum Entropy, which allows selection of the best from a series of different probability distributions that each one expresses the current state of knowledge. This informs which one is the preferred choice with the largest entropy (Aggarwal, 2015; Aggarwal & Reddy, 2014).

- **Support Vector Machines (SVM):** This can be applied to both classification and regression. Support vectors are data points nearest to the "hyperplane" (Aggarwal, 2015; Aggarwal & Reddy, 2014; Lantz, 2015). SVM tries to identify the "hyperplane" that best divides a dataset into two classes. A hyperplane is a 'line' that linearly separates and classifies a set of data. If the data points are further away from the hyperplane there can be greater confidence in a correct classification having been made.

The use of a supervised approach can help model each tweet as a vector of sentiment features. In addition, the datasets will use the manually annotated tweets for training and validation. When the feature vectors from all tweets have been extracted, they will be used together alongside the manually annotated sentiment labels as input for supervised learning algorithms. Several learning algorithms will be applied to fulfil this task e.g. SVM, decision trees and naïve Bayes. The results of the learned function will be applied to infer automatically the sentiment label for some unseen tweets.

The sentiment analysis outcome from the dictionaries to the machine learning approach will be evaluated with a series of techniques, which is explore in section 5.10.

5.10 Evaluation methods

It is important to measure the effectiveness of the outcome as classification algorithms have varying strengths and weaknesses, testing them will help distinguish among the learners (Bali and Sarkar, 2016; Lantz, 2015). Testing the result is imperative to forecast how a learner performs on future data. It is important to measure the accuracy of the results rather than just accepting it as the right classification. The classification result (label) of each sentiment dictionary was used as votes in the majority voting of all dictionaries (Bali and Sarkar, 2016; Lantz, 2015). The results of each lexicon will be totalled up in each sentiment category, where one can identify the majority category for each tweet. If there is a draw between the sentiment category, then a flip of the coin will decide which one is the winner.

Another method was employed to test each lexicon dictionary's results against the manually classified tweets to determine the accuracy of the results. Some possible measures of accuracy here are Precision, Recall and F1 (Bali and Sarkar, 2016; Lantz, 2015). Precision indicates what number of instances are relevant from the data e.g. defining the proportion of positive examples being truly positive. Precision is a portion of relevant positive/negative/neutral retrieved from the total retrieved (Bali and Sarkar, 2016; Lantz, 2015).

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Figure 5.1 Precision formula (Lantz, 2015)

Recall determines the number of elements that have been retrieved over the total number of relevant instances (Bali and Sarkar, 2016; Lantz, 2015). This is defined as the number of true positives over the total number of positives.

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Figure 5.2 Recall formula (Lantz, 2015)

Both precision (sensitivity) and recall (specificity) are based on the understanding and measure of relevance (Bali and Sarkar, 2016; Lantz, 2015). It can be difficult to build a model that has both high precision and recall, as it can be easy to obtain high precision if targeting easy to classify examples, but gives no indication if all the relevant positive/negative/neutral sentiment were retrieved (Bali and Sarkar, 2016; Lantz, 2015). High recall can follow a similar example, as the model can be overly aggressive in identifying positive cases, but provides no indication of how many retrieved documents are irrelevant (Bali and Sarkar, 2016; Lantz, 2015). Additionally, if precision

or recall is higher than the other, then it is important to test different models to find a good combination of precision and recall.

Precision and recall can be combined into F1 or F-Measure, which measures the accuracy of the classification as a whole (Bali and Sarkar, 2016; Lantz, 2015). F1 takes account of both precision and recall (Bali and Sarkar, 2016; Lantz, 2015). F1 is the harmonic mean of both precision and recall, where the F1 score of 1 is perfect precision and recall and 0 is the worst score with either no precision or no recall. F-Measure is calculated using the formula:

$$\text{F-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{recall} + \text{precision}} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}$$

Figure 5.3 F-measure formula (Lantz, 2015)

F-measure can describe the model's performance with a singular number enabling comparisons across several models against one another (Bali and Sarkar, 2016; Lantz, 2015). F1 can apply different weights to calculate the F-score for precision and recall, but it may be difficult to assign appropriate weights (Bali and Sarkar, 2016; Lantz, 2015). This could produce a positive or negative result, depending if the weight allocated is suitable for the context. Therefore, it is important to use these different measures to consider the models strengths and weaknesses (Bali and Sarkar, 2016; Lantz, 2015).

F-measure is useful to measure the performance of text classification in a way that is informative and more useful than classification accuracy (Athar, 2014; Gate, 2019; Zhang, Wang, Zhao, 2015). This is due to the established occurrence of class imbalance between positive/ negative/ neutral sentiment classification. When there are multiple classes present in a document collection, then the single aggregate F-measure is used that combines F1 scores from each class (Athar, 2014; Gate, 2019; Zhang, Wang, Zhao, 2015). Multi-class text classification performance is measured on the effectiveness based on macro-averaged and micro-averaged of F-measure scores (Athar, 2014; Gate, 2019; Zhang, Wang, Zhao, 2015). Macro averaging calculates precision, recall and f-measure on a per document basis, and then averages the results. Micro averaging treats the corpus as one large document, so calculates the average of the F1 scores over classes (Athar, 2014; Gate, 2019; Zhang, Wang, Zhao, 2015). The difference between these two methods are that the micro average provides equal weight to “each per sentiment classification decision,” thus making it dominated by large classes, while the macro average provides equal weight to each class (Athar, 2014; Gate, 2019; Zhang, Wang, Zhao, 2015). These indicators should not be a way to determine how reliable a classifier will be for future performance on unseen data (Athar, 2014; Gate, 2019; Zhang, Wang, Zhao, 2015). The average of F1 scores reflects on the sentiment classifier’s performance based on its given test data. If the micro average is lower than the macro average, there might be poor performance on the larger classes and, conversely, if macro average is lower than the micro average, then

there may be poor metric performance on the smaller classes (Athar, 2014; Gate, 2019; Zhang, Wang, Zhao, 2015).

The evaluation techniques explored for sentiment analysis can help understand how conclusive the results of any sentiment classification results. The next section of 5.11 focuses on identifying significant change points during the events. Additionally, the strongest results from the machine learning outcomes will be used in the change point process. Change points may help inform the police when it may be the appropriate time to adapt to changing situations at an event and/or whether to intervene online to maintain the peace as well.

5.11 Change-Point Detection

Change Point Detection (CPD) focuses on sequential detection of a change point by observing the process. This tries to identify times when the probability distribution of a time series changes (Isaac Newton Institute, 2017). The process tends to model measuring the quality of continuous process, where a change point may be identified in transition (Aminikhanghahi & Cook, 2016). This might indicate a deterioration or improvement in quality that is detected and eventually corrected (Aminikhanghahi & Cook, 2016). Change point analysis tries to detect anomalies of behaviour within the data. Detection of change points aids in modelling and prediction of time series and can be found in a broad range of applications that will usually present a variety of different problems (Aminikhanghahi & Cook, 2016), for instance, climate change detection and human activity analysis. These sequences of measurements over time describe the behaviour of systems, which can change due to external events and/or internal systematic changes in dynamics/distribution (Aminikhanghahi & Cook, 2016).

We will consider the application of such analysis to Twitter sentiment data to attempt to detect any change in the statistical mean. This will help to identify when a change has occurred as to pinpoint when a change has occurred may help attempt to identify its cause and predict future change (Aminikhanghahi & Cook, 2016). There are several questions that may help to understand the points of change (Kass-Hout & Xu, 2017; Killick, 2014; Killick & Eckley, 2014):

- Has a change occurred? If yes, where is the change?
- What is the difference between the pre and post change data?
- What is the probability that a change has occurred?
- How certain are we of the changepoint location?
- How many changes have occurred?
- Why has there been a change?

The project will illustrate how to apply change point analysis techniques in practice, through a series of use cases when approaching the four case studies, such as

identification of peak activity changes. The application of change point detection techniques for temporal analysis of social media data is nascent (Lansdall-Welfare, Dzogang & Cristianini, 2018), and it is especially difficult to locate papers on this area for public order events. Several papers have been identified based on a series of different events. Lansdall-Welfare, Dzogang & Cristianini (2018) studied online UK public mood in the days before and after the Brexit referendum and explained the variability of emotions with multiple change point analysis. This resulted in understanding that including other sources of variation can reduce unexplained movements by considering a grouping of both GBP/ EUR exchange rate and public mood (Lansdall-Welfare, Dzogang & Cristianini, 2018). This showed that positive sentiment had a positive correlation with the exchange rate, while a stronger anti-correlation was found for negative sentiment expressed in anger and sadness measured via Twitter (Lansdall-Welfare, Dzogang & Cristianini, 2018). This helps us to see links between forming of opinion and affective experiences. The monitoring of social media and traditional communications can provide *"insight into how events and policies influence public attitudes."* (Lansdall-Welfare, Dzogang & Cristianini, 2018, pg7).

Singh, Roy & Gangopadhyay (2018) suggest that data analytics on Twitter can help the disaster and emergency services to feedback to emergency responders and local authorities. An aspect of their research utilises change point analysis to process, uncover and infer the spatiotemporal sentiment of users based on the 2017 Las Vegas shooting. This research analysed sentiment polarity, but further improvement is required to understand the detailed emotions of the public in an event of crisis. To that end, the researchers focused on 8 different types of emotions to gain a greater insight and make the responders' event handling more emotionally aware. Additionally, Tasoulis, Vrahatis, Georgakopoulos & Plagianakos (2018) analysed real time sentiment change detection of streaming data using a cumulative sum (CUSUM) algorithm based on Brexit news topic. This means ensuring the methodology does not require an off-line phase or training. There was a focus on discovering propaganda efforts and spreading of fake news in early stages, alongside identifying sentiment changes of hashtags. The results focused on a moving average that moved from a positive to negative polarity vice versa. The overall direction of the moving average highlighted a slight negative sentiment. Furthermore, Goutte et al. (2018) detects changes with an online stream of tweets that are pre-processed and relies on linguistically relevant time series to run multivariate change point detection algorithm. The focus was on the 2016 Football European Championships, and this was then used as a benchmark to detect approximately half of the significant game play in a football game.

CPD algorithms are classified as "offline" or "online". Offline algorithms consider the whole dataset once collected and retrospectively seek to identify where the change arose (Aminikhanghahi & Cook, 2016). The purpose of doing this is to establish all

sequence change points in a batch mode. Meanwhile, online algorithms use streaming data near real-time to process and monitor every data point (Aminikhanghahi & Cook, 2016). The goal is to detect a change point after it directly occurs, ideally before the next data point arrives (Aminikhanghahi & Cook, 2016).

In our case, data collection has been completed already, so an offline approach will be used to process the data with the primary aim to accurately detect changes in sentiment. The pre-processing of data will be in accordance with Goutte et al. (2018), but we instead analyse the sentiment of a sample of tweets with lexicon dictionaries.

In section 5.12, the initial findings of each case study will provide background information on each demonstration, alongside linking it to any other previous studies relevant to these specific events.

5.12 Initial findings of each case study

The initial findings of each case study will be explored by analysing the timelines, visualising terms and lexical density to gain insight for each demonstration.

5.12.1 Timeline of events

5.12.1.1 Both MMM2015 and MMM2016

The initial findings from MMM 2015 and MMM 2016 have been analysed without pre-processing the data. Figure 5.4 displays a timeline of the 2015 event, which has a constant flow of tweets from 3rd to the 4th of November. However, the 5th of November 2015 MMM, there is a spike with a decline by less than half of the tweets posted on the 6th of November. On 7th November there is a further decline, with a slight drop on the 8th. There are 131,451 tweets classified as retweeted, but other retweets may go undetected as outlined in section 5.5. MMM2015 hashtags highly used are as follows:

- MillionMaskMarch (110426), Anonymous (65671), MMM2015 (44313)
- MMM (15042), Nov5th (9209), OPKKK (6123), London (4831)
- FreeAnons (3797), HoodsOff (2824) and KKK (2656)

The top three hashtags are already known as these are the search terms used to extract the data from Twitter. Some other hashtags, such as OPKKK, FreeAnons and KKK are more concentrated in other countries, mainly America, which is determined by location, hashtag or keyword specified in the tweets.

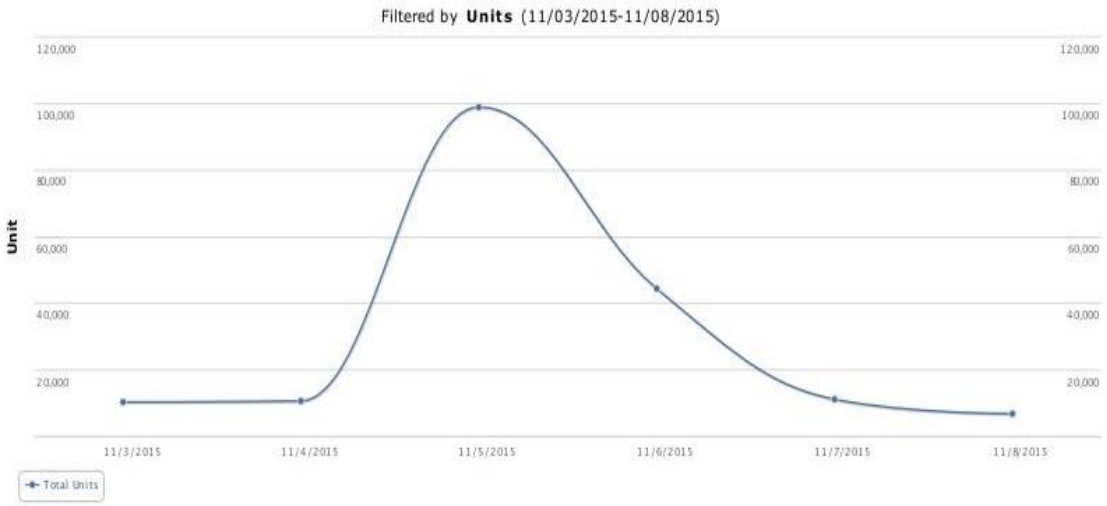

Figure 5.4 Tweets by Day MMM 2015

In Figure 5.5, MMM 2016 demonstration displays a similar trend as shown in Figure 5.4, but the main difference between the two marches is the total number of tweets, which is outlined as follows: -

- On 03/11/16 there was 6,978 (12am), on 04/11/16 7805 tweets (12am), with 56,190 on 05/11/16 (12am), to 23,541 on 06/11/16 (12am), on the 07/11/16 at 7498 and finally 6534 tweets on 08/11/16 (12am).
- The highest peak in tweets on the 5th of November in 2015 was 98,787 (12am) compared to 56,190 (12am) in 2016, which shows a considerable decline in participation on specific day. There are 79,478 tweets retweeted and others may have gone undetected.

The number of tweets for the event shows a large decline compared with 2015 MMM. The 2016 MMM hashtags highly used are as follows: -

- Anonymous (53466), MillionMaskMarch (39808), MMM2016 (27428)
- MMM (6513), WikiLeaks (5136), London (3601), PodestaEmails31 (3001)
- MMMLiveOnThe5 (2978), Nov5th (2205) and MMMlondon (1929).

The top three are the same hashtags used to extract the data similar to 2015 MMM, but the remaining show a higher association with the event as MMM used multiple times in the hashtags and two are associated with London. 'PodestaEmails31' is unusual, therefore, it was investigated and is not based on this specific event. This seemed to be more linked with another area of the Anonymous movement.

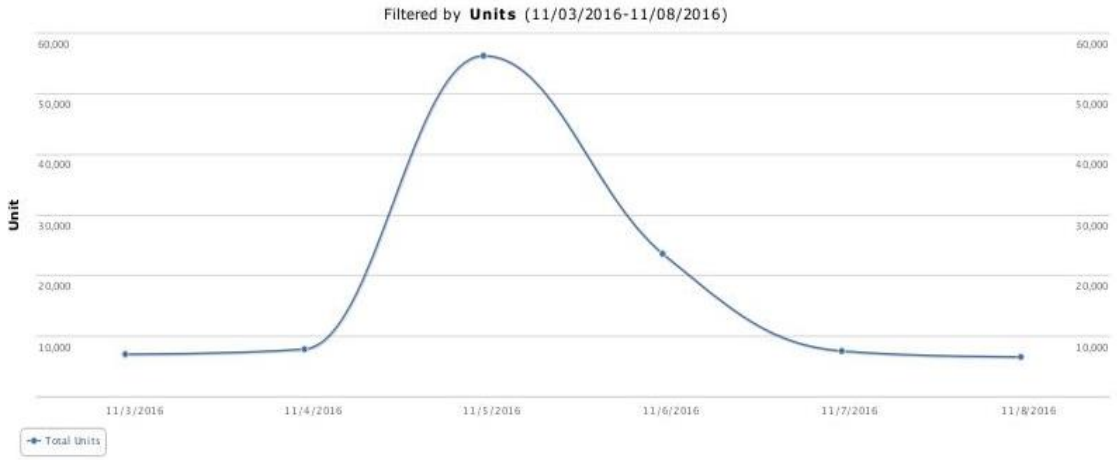

Figure 5.5 Tweets by Day MMM 2016

In Figure 5.6, MMM 2015 data is presented by day and hour, which provides a clearer breakdown where number of tweets has peaked at a specific time rather than the total count for each day. Figure 5.6 is described as follows: -

- On 03/11/15 at 12am 813 tweets, are followed by a rise on 05/11/15 at 12 noon to 1156, but by at 8pm it rose to 13,249 tweets and dropped to 6319 at 12am. The lowest number of tweets is 165 tweets at 08/11/15 at 7am.

- The tweets significantly rose on the day of the event, as there was 1,556 by midday to 9,727 tweets by 6pm and 11,952 by 7pm. The highest peak was at 8pm with 13,249, but kept decreasing from 9pm (11,264) and by 11pm was 7,486, which showed a significant decline. This was the point where most dispersed from the event. The decline of tweets kept on reducing after 11pm, by morning it fell to 1,423 on 06/11/15 at 6am and reduced to hundreds of tweets in following days.

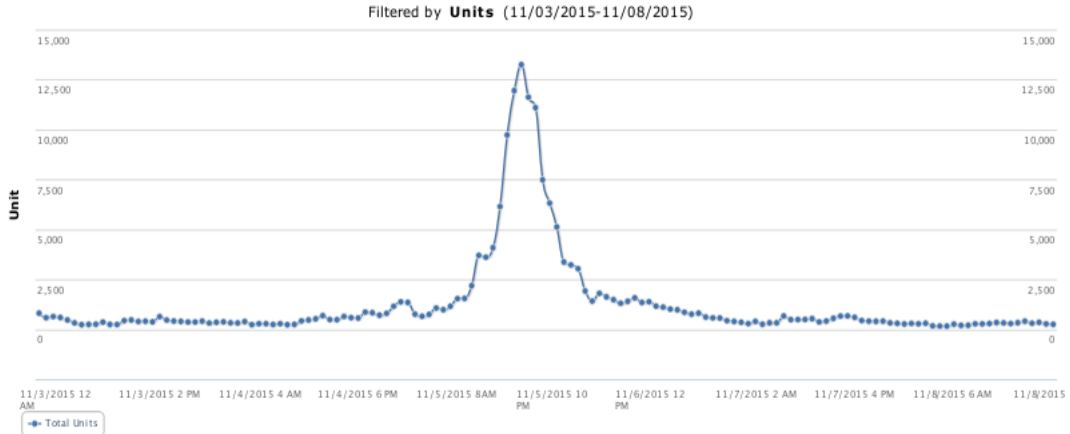

Figure 5.6 Tweets by Hour MMM 2015

In comparison with 2015 MMM, Figure 5.7 shows there were much less tweets over the same frame, as 2015 MMM saw 181,711 tweets, but MMM2016 has 108,546 tweets. In between 6th to 8th November displays a similar number of tweets. Additionally, on 5th November at 12 noon saw 56 less tweets posted and again the highest peak was at 8pm, but with 6911 less tweets posted, which was half the amount. Furthermore, there was a large increase from 1100 tweets at 12 noon to 6122

at 6pm, followed by 6134 at 7pm to 6258 at 8pm. In 2016 after the highest peak saw a rapid decline as the event dissipated at 9pm rather than around 11pm in 2015. In the days after the event, it followed a similar trend to 2015, where it constantly declined in tweets over time.

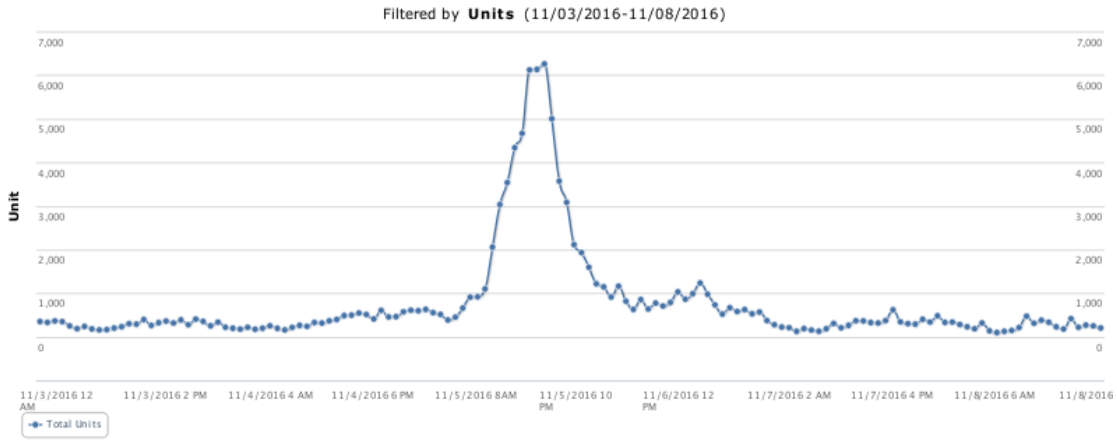

Figure 5.7 Tweets by Hour MMM 2016

The results show significant change occurs on the day of each MMM event. Both follow a similar trend where the event discussion tends to rise day before, peak on day of demonstration, and then reduce dramatically over a couple of days.

5.12.1.2 Anti-Austerity

In Figure 5.8, the Anti-Austerity 2016 demonstration shows on 13/04/16 there were 29,336 at 12am, which may indicate there is a greater public interest on this topic than MMM events. On 14/04/16 at 12am 29,180 tweets were posted, which rises to 29,284 on 15/04/16 (12am). This demonstrates a stronger affinity with the subject as the number of tweets stay at a similar volume.

On the day of the event (16/04/16) it significantly rose to 98,915 tweets being posted, which again similarly declined the time after the event with 35,813 on 17/04/16 at 12am and finally 27,888 tweets on 18/04/16 at 12am. The highest peak is on the day of the event, but shows a considerable decline in participation after the day of the event similarly as depicted in both MMM events. There are 136,820 retweeted, but other retweets may be undetected. The hashtags highly used are as follows:

- London (154540), 4Demands (86508), jobs (10960), UK (8096)
- resigncameron (4713), Paris (4437), TFL (4383)
- CameronMustGo (3540), USA (3370) and germany (2979)

Both ‘London’ and ‘4Demands’ are chosen for extraction show a strong association with the topic of the event apart from ‘Paris’, ‘USA’ and ‘Germany’ which refer to austerity within their respective countries. Overall, these hashtags show a stronger association to the event than both MMM event.

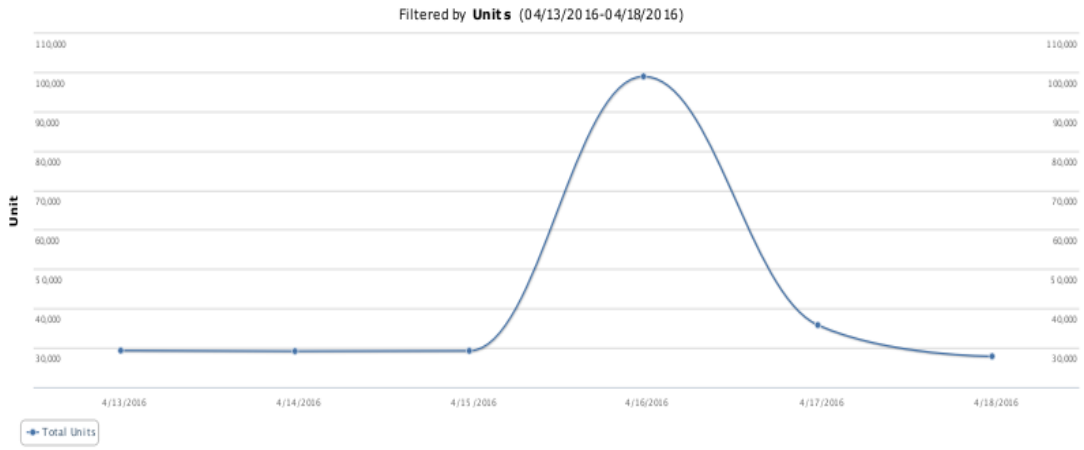

Figure 5.8 Tweets by Day Anti-Austerity

In Figure 5.9, there are 1,941 tweets at 1pm on 13/04/16, 14/04/16 4pm 1,830, 15/04/16 11am 1,923. On the day of the event there are 3,575 tweets posted at 9am and rose sharply to 10,418 tweets per hour, the highest peak. After 2pm the number of tweets declined less rapidly as the other events due to the finish time being earlier in the day than MMM events late at night. The tweets posted declined at 5pm 8,369 to 6pm 5,880 and by 12am it reduced to 1,350 and the lowest peak is on 18/04/16 with 401 tweets at 3am.

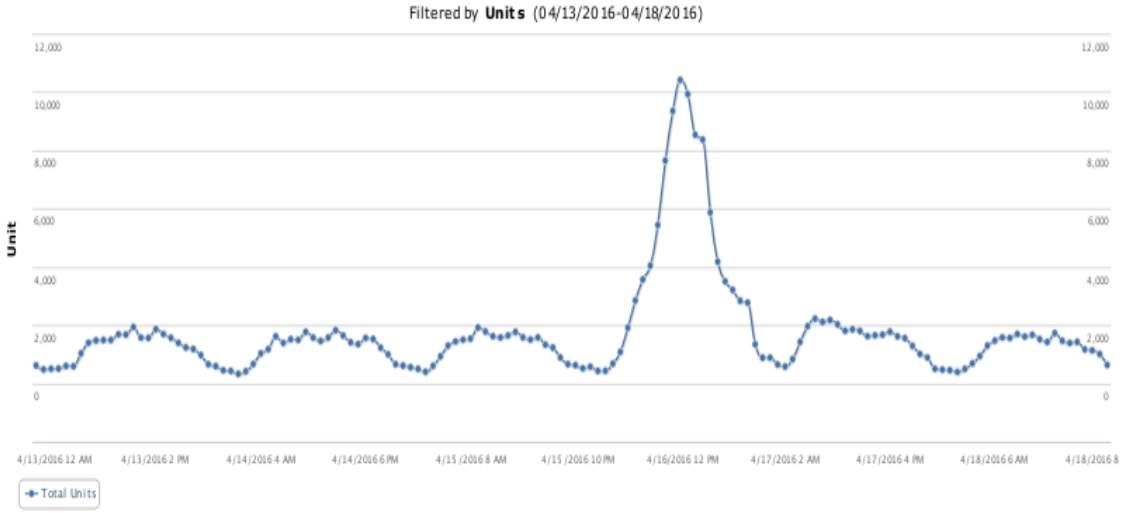

Figure 5.9 Tweets by Hour Anti-Austerity

5.12.1.3 Dover 2016

In Figure 5.10, three days before the Dover demonstration on 30th of January, there were 1,151 tweets posted on 27/01/16 by 12am and which lowered to 1,009 on 28/01/16 by 12am. This rose to 1,866 tweets on 29/01/16 by 12am, which largely rose to 14,949 tweets on 30/01/16 by 12am. Similarly as the other three case studies, the day after the event saw a drop, which was from 3,723 tweets 31/01/16 by 12am to 2,333 on 01/02/16 by 12am. There are 17,826 retweeted and other retweets may be undetected. The Dover 2016 hashtags highly used are as follows:

- dover (14483), antifa (9922), antireport (1206), athens (1136)
- Greece (849), RefugeesWelcome (521), antinazigr (488)
- jobs (381), Nopasaran (320) and refugees (295)

The top two hashtags are chosen for extraction, but these hashtags are mostly related to other countries or other topics than the event in Dover when compared with the other case studies.

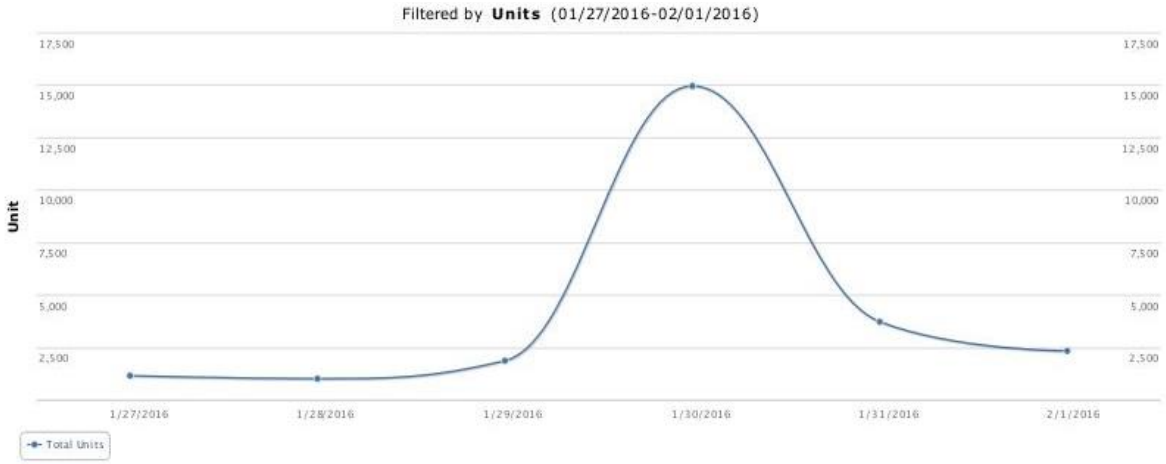

Figure 5.10 Tweets by Day Dover

In Figure 5.11, 20 tweets were posted on 27/01/16 20 at 12am and rose to 141 tweets at 10am at its highest peak that day. This did not rise any higher until the 29/01/16 at 4pm where 171 tweets were posted. On the day of the event, at 9am 106 tweets were posted, which rose steadily to 1,781 per hour at 4pm, by 8pm dropped to 872 tweets and by midnight to 252 tweets. The lowest peak is 21 tweets on the 01/02/16 at 1am. The rest of the tweets in the day after the event are either less than 100 or in the couple of hundreds of tweets. This dataset in comparison with the other case studies has considerably less volume of tweets.

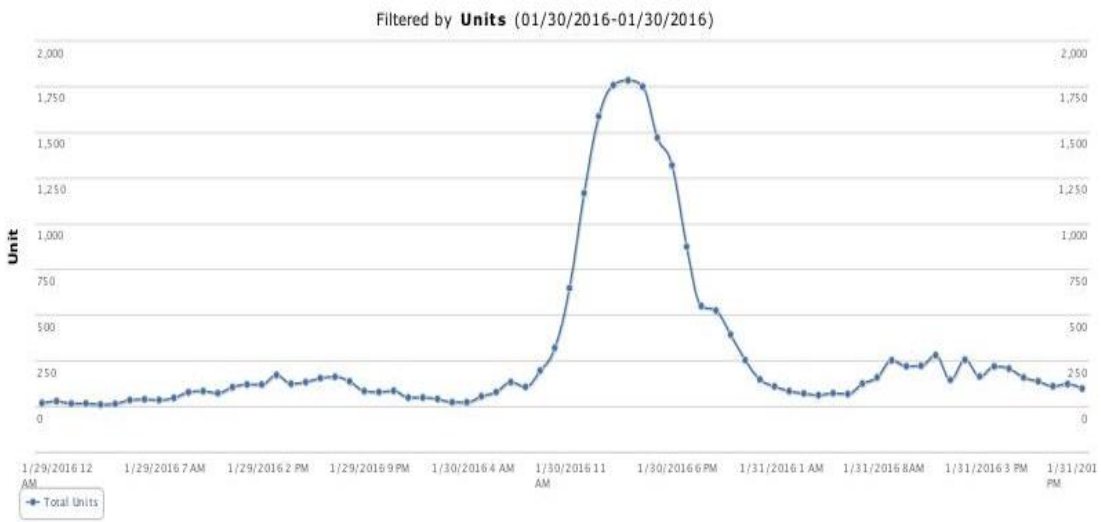

Figure 5.11 Tweets by Hour Dover

The tweets posted by at each of the events are predominately by mobile devices, such as “Twitter for Web Client”, “Twitter for iPhone”, “Twitter for Android” or “Twitter for iPad”. The source of the tweets in-terms of the location of these devices are mostly unknown. There are ways to identify the location of the individuals by geo-coordinates, user location and location keywords/hashtags. Additionally, the actual location can be accurately determined by triangulating all the mobile devices in the area. This is difficult to access without relevant authorities and telecommunication agreement.

5.12.2 Word cloud analysis

In the pilot study, word clouds proved useful to provide a firmer understanding of the event. In this iteration of word clouds for each event, stop words and other unique words have been removed along with some of the top TF-IDF terms removed with no semantic value from the corpus, thus reduces the dimensionality of the input space. In the first word cloud Term Frequency (TF) is applied with stop words and other unique words removed. In the second is the same with the inclusion of TF-IDF which is a sum of TF-IDF of each word across all tweets. These two will be compared to identify which approach is strongest in semantic value to describe the events.

In 2015 MMM, both Figure 5.12 and Figure 5.13 word clouds have been effective in describing the event with these keywords, such as topic of event based on ‘anti-capitalism’, demonstration has been ‘peaceful’ and ‘violent’, ‘fireworks’ let off, may be a ‘car’ ‘fire’, ‘horses’ and ‘officers’ on the ‘streets’ where ‘arrests’ were made at the event. Additionally, Figure 5.13 with the use of TF-IDF has provided a similar picture, but has provided additional information that the event may be streamed and clashes occurred at the event. Both word clouds have shown a good selection of words, but could be improved by adding more stop words to list, such as ‘going’, ‘tonight’, ‘underway’ and ‘watch’, which describe little about the event.

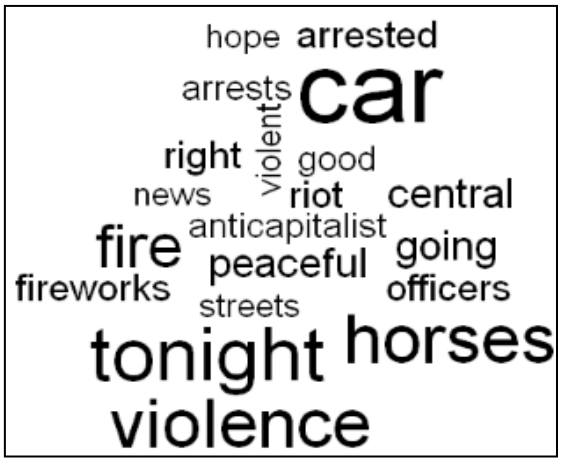

Figure 5.12 MMM 2015 Word Cloud

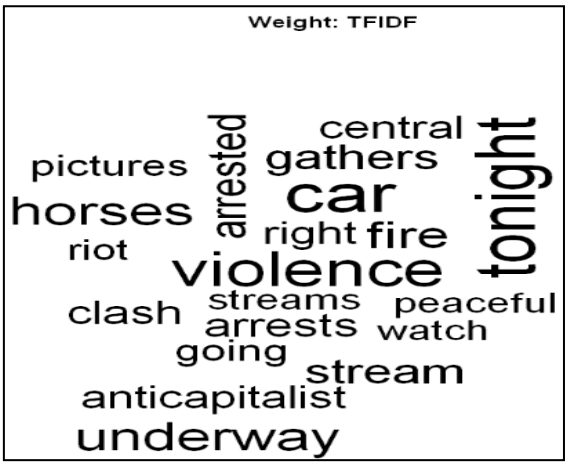

Figure 5.13 MMM 2015 Word Cloud TF-IDF

In 2016 MMM both Figure 5.14 and Figure 5.15 word clouds have provided an insight into the event. Figure 5.14 shows the event is largely peaceful with elements of being 'happy' and be 'safe', but with use of 'fireworks' and 'violence' there have been some 'arrests' in the evening (e.g. 'tonight') demonstration. Figure 5.15 displays largely the same picture, but with added emphasis on 'happy' and 'arrests' and other words that indicate the event may be in a 'central' location.

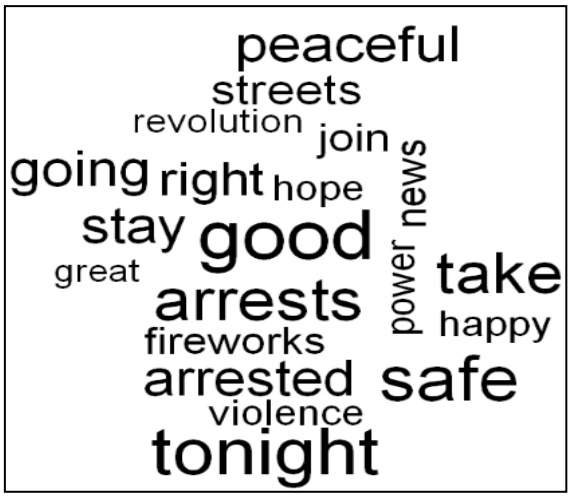

Figure 5.14 MMM 2016 Word Cloud

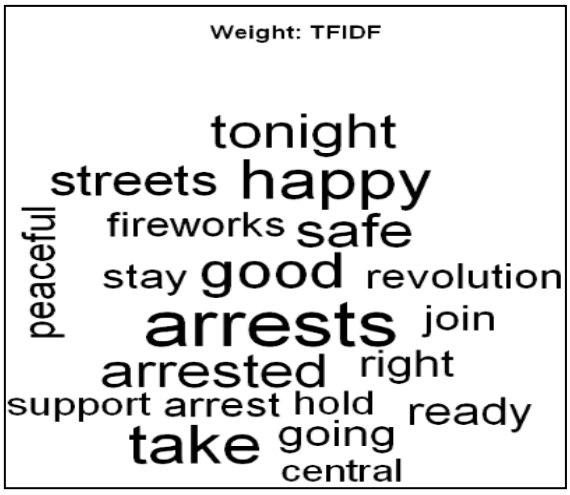

Figure 5.15 MMM 2016 Word Cloud TF-IDF

In both Figure 5.16 and Figure 5.17 word clouds of the 2016 Anti-Austerity demonstration show the main focal point is government making ‘cuts’, ‘homes’, ‘education’ and ‘NHS’. Additionally, the public is showing ‘solidarity’ and ‘support’ for anti-austerity demonstration, but have outlined a ‘reporting coverage’ as a problem. Most of the case studies show a higher level of negativity, which can be identified within each of the word clouds, such as ‘arrests’, ‘clashes’, ‘violence’ and some profanity.

Figure 5.16 AA 2016 Word Cloud

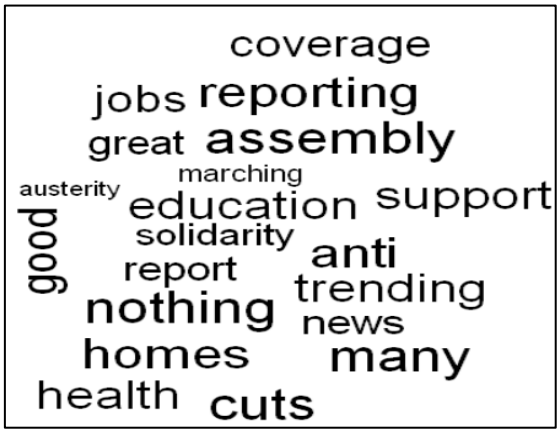

Figure 5.17 AA 2016 Word Cloud TF-IDF

In both Figure 5.18 and Figure 5.19, the 2016 Dover demonstration was more repetitive in nature with certain variations of the same word appearing, such as ‘antifacists’ and ‘antifascist’ or ‘fascism’, ‘fascist’ and ‘fascists’. The inclusion of these added words does not provide much value to the describe the event. In the future, stemming may be employed to ascertain whether this would mean other important words are included to provide a greater insight to the event. Additionally, this dataset is smaller than the other datasets by thousands of tweets. This difference in dataset size may provide lesser variation of words. The event shows there is a ‘fascist’ and ‘anti-fascist’ ‘groups’, where there may be ‘violence’ between the two groups. In Figure 5.18, the ‘fascists’ are being called ‘nazis’, ‘scum’ and ‘thugs’, but on the other side there ‘solidarity’ towards ‘immigration’. Additionally, in Figure 5.19, it emphasises there are ‘clashes’ and something funny has occurred with the inclusion of ‘lol’.

Figure 5.18 Dover 2016 Word Cloud

Figure 5.19 Dover 2016 Word Cloud TF-IDF

5.12.3 Lexical Density

Russell (2019) and Inuwa-Dutse, Shehu Bello, and Korkontzelos (2018) have used lexical density as a quantitative measure for the range of vocabulary for an individual or group to help understand the language use and complexity of the text for a subject matter. The lexical diversity is number of unique tokens divided by number of total tokens. Lexical density has been applied to each of the case studies.

In Figure 5.20, the 2015 Million Man Mask demonstration (total 3296 tweets) at the highest peak is 14 words for approximately 270 tweets, whereas the lowest is at 1 word and 28 words for 2 tweets. Most are above 100 tweets from 6 words until 23 words. In Figure 5.21, the highest peak is at both 14 and 15 unique words per 300 tweets, whereas the lowest is 27 words for 1 tweet. Most are above 100 tweets from 6 unique words until 23 words. The lexical diversity result is 0.25 which is very low.

[Intentionally Left Blank]

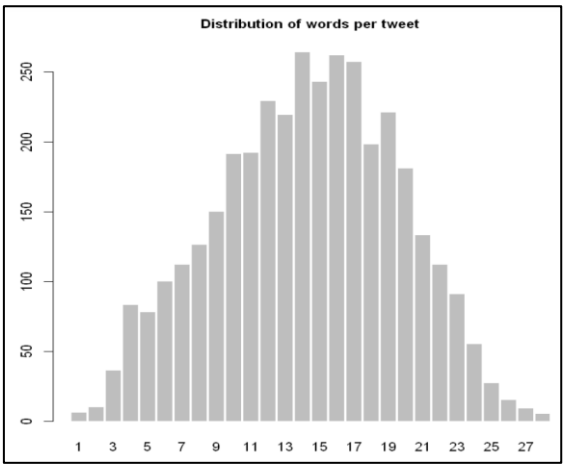

Figure 5.20 2015 MMM Distribution of words per tweet

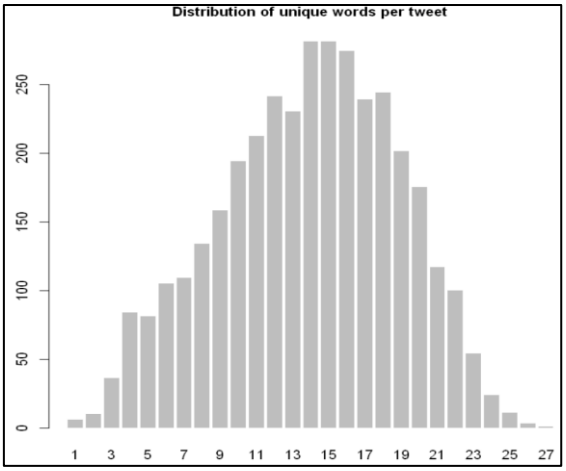

Figure 5.21 2015 MMM Distribution of unique words per tweet

In Figure 5.22, for the 2016 Million Man Mask demonstration (total 3356 tweets), the distribution of words per tweet is at its highest peak at 16 words per roughly 250 tweets, whereas the lowest peak is at both 1 and 29 with 1 per tweet. Most tweet counts above 100 are from 4 to 22 words, but at 23 words see a decline below 100 tweets. In Figure 5.23, the highest peak is at 16 unique words per approximately 250 tweets whereas the lowest is 1 unique word for 1 tweet, with 27 unique words is for 2 tweets. Most above 100 tweets are from 4 until 22 unique words. The lexical diversity is 0.27.

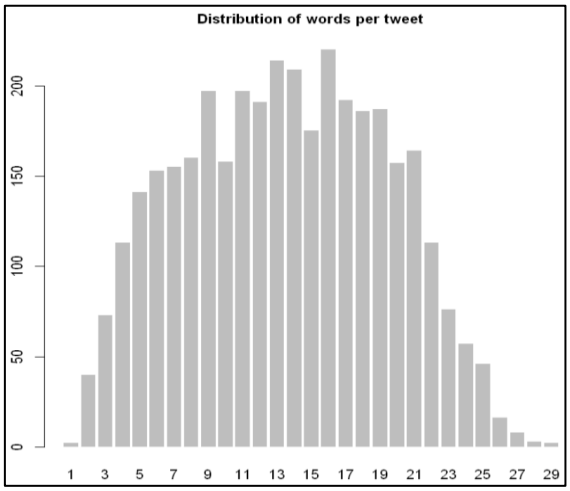

Figure 5.22 2016 MMM Distribution of words per tweet

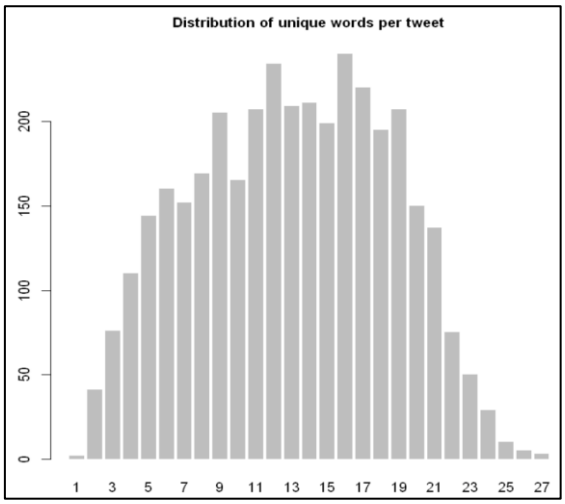

Figure 5.23 2016 MMM Distribution of unique words per tweet

In Figure 5.24 of the 2016 Dover demonstration (total 2830 tweets), the distribution of words per tweet ranges from approximately 200 tweets contains 14 words per tweet, to the lowest being 28 words for 1 tweet. Most tweets counts over 100 have 4 to 22 words, but again at 23 words on-wards to 27 see a decline in tweets below 100. In Figure 5.25, the highest peak is 18 unique words for 220 tweets and most tweets above 100 have 2 or more unique words ranging up to 26. The lowest peak is 1 unique word for 1 tweet, and the other remaining low ones are 23-27 unique words. Moreover, others tweets range somewhat below 100 with 3-6 unique words. The lexical diversity is 0.27.

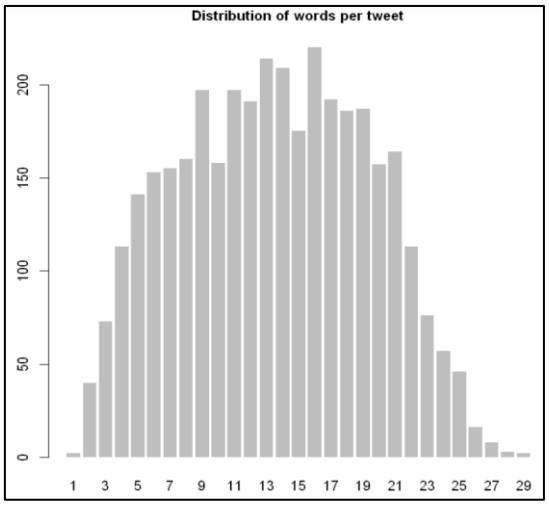

Figure 5.24 2016 Dover Distribution of words per tweet

[Intentionally Left Blank]

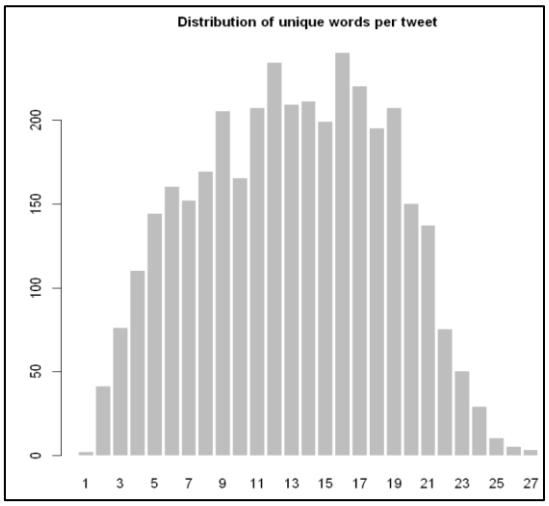

Figure 5.25 Dover Distribution of unique words per tweet

In Figure 5.26, 2016 Anti-Austerity demonstration (total 5446 tweets), the highest peak in tweets of approximately 350 tweets for 8 words, with the lowest peak being 28 words per 1 tweet. Most above 100 tweets are 2 words and above until 23 words, where it goes below 100 tweets. In Figure 5.27, the highest peak is at 8 words for 350 tweets, with the lowest being at 26 and 27 unique words per 1 tweet. Most are above 100 tweets that range from 2 to 21 unique words. The lexical diversity is 0.11.

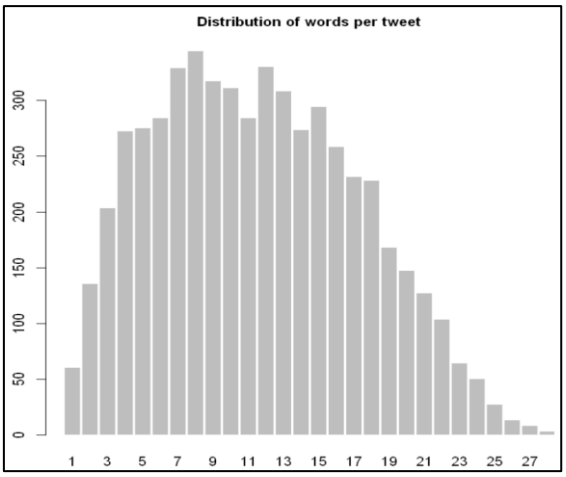

Figure 5.26 2016 AA Distribution of words per tweet

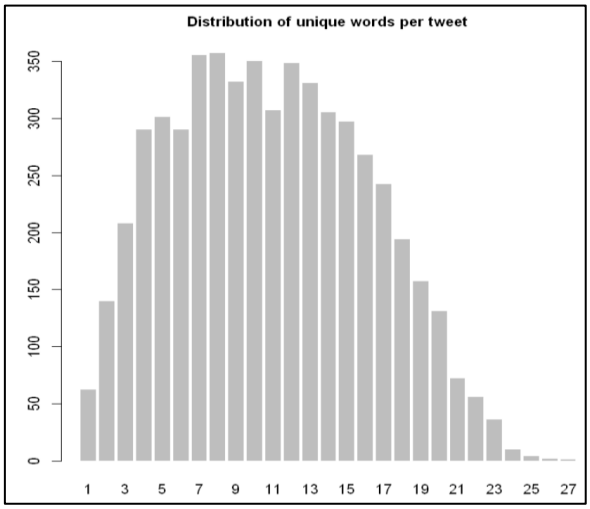

Figure 5.27 2016 AA Distribution of unique words per tweet

The highest distribution of words per tweet with the largest number of tweets is word count of 14 and 16 for the first 3 datasets results with the exception with Anti-Austerity on 8 words. The highest average is 15.33 based on the first three datasets, but with Anti-Austerity, it reduces to 13.5. Each dataset mainly has 1 tweet with either 1 or 28/29 words, except for Anti-Austerity with 28. The lowest average is either 1 or 28.5 words for 1 tweet.

The highest distribution of unique words per tweet graphs show similar results to words per tweet with the first three datasets on 14/15, 16, 16 and Anti-Austerity on 8. The only difference is for 2015 MMM where it has two bars on the same highest number of tweets being on 14 and 15, but averaged is 14.5. If averaged with the next two it is 15.5 words, and the final one with 8 is 13.63. The average of the total for words and unique words has difference of 0.30. Lowest unique words vary from 27, 1, 1 and 26.5, but there are only a couple of tweets different for second and third, which would have been similar to first and last. The lowest average is 13.88 for 1 tweet.

The highest lexical diversity is for 2016 MMM and Dover on 0.27 and lowest is 0.11 for Anti-Austerity. The lower lexical density may indicate a lesser understanding of subject matter, whereas on a higher scale would be the opposite. However, the lexical diversity may be lower for tweets as the users in 2015 and 2016 are limited to 140 characters per tweet. Despite low lexical richness, there is value in the information. In future studies of lexical diversity for Twitter may be more useful indicator as Twitter raised to 280 characters per tweet.

5.13 Summary

The time series graphs above have provided a firm understanding of the pulse of discussion over time, which requires further investigation in the sentiment analysis phase. The sentiment analysis results are less reliable with the 'sentiment' package as there is high mis-classification. This Berkeley dictionary will be tested along with other dictionaries on transformed datasets for the case studies to identify if Berkeley dictionary is still an issue and how this compares with the other dictionaries results. Moreover, the focus will be polarity as a wide range of emotions provide a less accurate outcome. The devices with most tweets are from mobile, such as Apple iPhone/ iPad or Android, which confirms that there is a greater use of mobile devices.

The word cloud helped to describe what happened in the event retrospectively, which the police could use to identify the change of topic on live streams of data for events. This may help to gain a quicker insight into an event. Additionally, the 'word cloud' approach has shown insight into each of the cases and TF-IDF has shown some improvement to describe the event with a few more relevant words than TF. Furthermore, the inclusion of stop words in this instance has provided a wider view of the event. The word cloud can be compared with the time series results to identify if

there are any differences and whether it strengthens the position of the overall outcome. Most words with more than 100 tweets are between the range of 6-23 for both words and unique words distribution. Moreover, these results both have similar high and low peaks for both words and unique words. The patterns in the distribution of words, unique words and lexical diversity are consistently low in lexical diversity which is mainly due to limit of characters per tweet. This will not be examined any further as this has provided a conclusive result to the lexical diversity and will not provide anything different in comparison with other results.

In section 6 we will explore the sentiment analysis and change point analysis results, which may further support the background findings and/ or discover greater insights.

6 Sentiment Analysis and Change Point Analysis Results

The first step of the Figure 3.4 hybrid approach is data extraction which has already been collected and described in section 3.1.6.2. The second phase is coding the data based on relevant and irrelevant data for each event. The third stage is to cleanse the dataset to gain a broad understanding of the data through the analysis of the tweets' sentiment by applying each lexicon-based dictionary.

The implementation of the cleansing approach adopted in section 5.8 removed irrelevant text or symbols to improve the data for analysis except for the use of stop words. The standard stop-word list applied appeared at first to remove a few too many words, but after closer inspection the words it contained are of less importance, such as MMM and Million Mask March. The dictionaries applied are outlined in section 5.9. Both dictionary and machine learning sentiment analysis approach are in carried out using R and the results are contained in series of Microsoft Excel spreadsheets. The standardisation process was applied to specific lexicon dictionaries as described in section 5.9 and applied to the same tweets for each study. The next step of the process is to follow the evaluation as outlined in both sections 3.1.6.1 and 3.1.6.2 to determine the strength of both dictionary and machine learning methods in the hybrid approach. This will help answer the second research question on which dictionary is the best and third question on whether a dictionary method is more effective than machine learning approach.

6.1 First Stage: Data Extraction

The data has been already extracted as specified in section 5.12. The tweets require coding based on their relevance/irrelevance for each event in the second stage.

6.2 Second Stage: Coding the datasets

The collected data at first will be manually coded (relevant) tweets to build a list of keywords to identify relevant and irrelevant tweets for each event as specified in section 5.7. A small sample of the data will be manually coded for each dataset. The keywords listed will be used to automatically code (relevant) tweets from the dataset for each event. The relevant tweets have been both manually and automatically coded.

6.2.1 Automated Coded Data

The keywords created in the manual coding will be used to identify relevant and irrelevant tweets for each event. For each keyword that is relevant it will be scored with a +1 and any irrelevant will be -1 similar to a sentiment analysis process but this time on relevance rather than affection. The total number of tweets started with, and number of tweets processed are stated in Table 4 used for each dataset. This total number does not include retweets which are automatically removed from the datasets in the cleansing phase as outlined in section 5.8.

Table 4 shows the initial results of the classification of which tweets are relevant and irrelevant. The automated results show all occurrences are mostly between 20% and 30%. The tweets classed as zero were reviewed which showed a large proportion of these were incorrectly classified as zero. As a result, the proportion of relevant and irrelevant tweets was lower than expected. Consequently, the keywords lists was extended with new words to increase relevant and irrelevant categories.

Classification of relevance results						
Dataset	**Automated/ Manual Coding**	**-1**	**0**	**+1 ABOVE**	**Total Tweets**	**Total Percent Coded**
AA	Automated coding	12,587	86,385	14,624	113,596	12.87%
	Manually coded	73	1,912	3,461	5,446	63.55%
2016 MMM	Automated coding	3,386	19,946	6,500	29,832	21.79%
	Manually coded	21	1,682	1,653	3,356	49.26%
2015 MMM	Automated coding	3,906	34,061	12,293	50,260	24.46%
	Manually coded	8	635	2,653	3,296	80.49%
2016 Dover	Automated coding	532	4,646	2,027	7205	28.13%
	Manually coded	54	1,245	1,531	2830	54.10%

Table 4 Automated keyword list coding

The keywords list was extended by adding both the most frequently counted words and Frequency–Inverse Document Frequency (TF-IDF). As a result, of this change the number of relevant and irrelevant tweets increased with fewer being unclassified. Table 5 results shows that the manual coder seemingly codes correctly, so this would

suggest the proportion of relevant tweets is highest for MMM 2015, but all datasets have over 80% relevant. However, the automated process is still very poor, with it finding only half that proportion apart from AA where it is worse still and finds only 26.45%

Classification of relevance results – Extended key words list						
Dataset	**Automated/ Manual Coding**	**-1**	**0**	**+1 ABOVE**	**Total Tweets**	**Total Percent Coded**
AA	Automated coding	33,242	50,310	30,044	113,596	26.45%
	Manually coded	88	946	4,412	5,446	81%
2016 MMM	Automated coding	2,170	12,111	15,551	29,832	52.13%
	Manually coded	2	469	2,885	3,356	86%
2015 MMM	Automated coding	2,436	18,214	29,610	50,260	58.91%
	Manually coded	4	180	3,112	3,296	94.42%
2016 Dover	Automated coding	420	3577	3208	7205	44.53%
	Manually coded	22	524	2284	2830	80.71%

Table 5 Automated extended keyword list coding

The automated relevant data will be used for sentiment analysis process with the 19 dictionaries including the combined dictionary. The pre-processing of the data has left some tweets blank with no score which are removed from each dataset, as follows: -

- Anti-Austerity has had 81 removed out of 30,044, now 29,963 after reductions
- 2016 Dover has 34 removed out of the 3,208, now 3,174 remain after the reductions
- 2016 MMM has removed 60 out of 15,551, now 15,491 remain after the reductions
- 2015 MMM has removed 190 out of 29,610 now 29,420 remain after the reductions

In the above it shows a proportion of the tweets are selected from the four datasets, and can proceed to the third stage for pre-processing the data.

6.3 Third Stage: Data Pre-processing

Both manual and automated coded (relevant) tweets are pre-processed with data cleansing techniques applied which are specified in section 5.8. In section 6.4, the cleansed tweets for each dataset will be validated for its reliability.

6.4 Evaluation of Sentiment Analysis

A sample of each dataset will be evaluated to validate the dictionaries' reliability. Each tweet in the sample will be manually classified by a series of researchers to measure the reliability.

6.4.1 Manual classification

The project uses various algorithms to score up and then analyse various Twitter datasets. In order for these algorithms to work correctly, a sample of the Tweets must be manually scored for sentiment. The manual classification will be for each dataset, where a sample of 1500 tweets have been classified as positive, negative or neutral. The judgement of whether the tweet fitted into one those categories was based on subjectively analysing the individual tweet without the aid of any other details to remain objective. It is essential that the reliability of these scores be measured (Bobicev & Sokolova, 2017; Landis & Koch, 1977; Santos, Bernardini, Paes, 2021). This can be achieved by other researchers to re-classify the same sample of data to determine the degree of agreement among raters. This is essential for measuring the reliability of the classified data and also the reliability of the algorithms (Bobicev & Sokolova, 2017; Landis & Koch, 1977; Santos, Bernardini, Paes, 2021). The manual classification of each dataset will be rated by three manual raters, which are known as MR1 (Manual Rater (MR)), MR2, and MR3.

Inter-rater reliability (also known as inter-rater agreement) is a score of the consistency in ratings provided by the same individual across multiple instances to test the validity (Landis & Koch, 1977; Liu, 2020). The assessment of this is useful in the refinement of the application provided by human assessors, such as determining if a scale is appropriate for the measurement of a specific variable. If the raters are in disagreement, then the scale may be defective or the rater requires re-training, so the task is effectively carried out. There are a range of statistics that can be applied for the inter-rater agreement, such as Cohen's kappa and Krippendorff's alpha (Bobicev & Sokolova, 2017; Landis & Koch, 1977; Santos, Bernardini, Paes, 2021). The comparison of the raters will make use of these, which are listed below together with an indication of how they should be interpreted:

- Judgement for the estimated kappa about the extent of agreement is given by Landis & Koch (1977):
    - If kappa is less than 0, "No agreement",
    - if 0-0.2, "Slight agreement",
    - if 0.2-0.4, "Fair agreement",
    - if 0.4-0.6, "Moderate agreement",
    - if 0.6-0.8, "Substantial agreement",
    - if 0.8-1.0, "Almost perfect agreement".

- P-value tests whether the estimated kappa is not due to chance. A value of 0.05 or smaller would indicate it is unlikely to be due to chance, the smaller this p-value the more unlikely this is (Bobicev & Sokolova, 2017; Landis & Koch, 1977).

- Cohen's Alpha may be calculated, differently using different assumptions about the weighting applied, but for the purposes of this study this is not relevant (Bobicev & Sokolova, 2017; Landis & Koch, 1977).

- Krippendorff's Alpha values range from 0 to 1, where 0 is perfect disagreement and 1 is perfect agreement (Bobicev & Sokolova, 2017; Landis & Koch, 1977).

The manual raters MR1, MR2 and MR3 results inter-rater reliability will be explored in section 6.4.2.

6.4.2 Inter agreement results

Table 6 shows a summary of the results for the comparison between raters MR1 and MR2 for all four data sets. It can be seen that for both the MMM data sets and for Dover that the agreement is moderate, (Krippendorff's alpha is about 0.5 as are all the Cohen's kappas) and that the agreement is over 70%. The exception is for AA which shows only fair agreement (as shown in full within Table 78 in appendix 10.10). Additionally, the p-value for Cohen Kappas is 0, which means the results are statistically significant, thus the appraiser agreement is significantly varied from what could be achieved by chance for all four datasets and all versions of kappa. Both MMM 2015 and 2016, and Dover provide similar results, whilst AA does not which shows the agreement is lower.

[Intentionally Left Blank]

Summarised Inter Agreement Results				
Level agreement				
Sentiment	MMM 2015	MMM 2016	Dover 2016	Anti-Austerity 2016
Negative	494	366	942	185
Neutral	521	600	190	581
Positive	54	88	13	74
Disagree	431	446	355	660
Total	1500	1500	1500	1500
Proportion				
Negative	32.93	24.40	62.80	12.33
Neutral	34.73	40.00	12.67	38.73
Positive	3.60	5.87	0.87	4.93
Disagree	28.73	29.73	23.67	44.00
Total	100	100	100	100
Percentage agreement (Tolerance=0)				
%-agree =	70.3	71.3	76.3	56
Krippendorff's alpha				
Alpha	0.511	0.527	0.453	0.271
Cohen's Kappa for 2 Raters (Weights: equal)				
Kappa	0.516	0.515	0.448	0.302
z =	25.5	27.1	21.1	19.8
p-value =	0	0	0	0
Cohen's Kappa for 2 Raters (Weights: squared)				
Kappa	0.54	0.549	0.477	0.359
z =	22.2	22.6	20.2	17.3
p-value =	0	0	0	0
	Cohen's Kappa for 2 Raters (Weights: unweighted)			
Kappa	0.5	0.492	0.429	0.264
z =	24.3	25.3	20.1	16.9
p-value =	0	0	0	0

Table 6 Summarised Results for Inter Agreement for MR1 & MR2

In Table 82 to Table 85 (refer to these tables in appendix 10.10) results from MR1 and MR3 consistently showed high level of disagreement with each dataset with the proportion of agreement ranging from only 48.9% through to 22.1% and similarly low Kappa and Alpha values. As a result, these results will not be explored any further, as precision and recall phase will be poor. Table 86 to Table 89 (refer to appendix 10.10) compare MR2 and MR3 results, which are slightly more mixed since both MMM datasets show around 60% agreement, but this is lower for both Dover and AA. Therefore, these results show a lower level of agreement compared to both MR1 with MR2, thus the precision and recall phase will be poor if using MR3. In Table 90 (refer to this table in appendix 10.10) from MR1, MR2 and MR3 agreement for all datasets is very poor, where most are 40% in agreement and further below is 2016 Dover on 18% agreement. This very poor agreement between MR1, MR2 and MR3 would not produce enough data for generalisation for the machine learning process, so there would be no point in taking any further action with this data as no greater insight can be retrieved.

In summary, the results from the inter-agreement have shown MR1 and MR2 to be the most reliable, therefore, MR3 will not be analysed any further in the next phase. The pre-processing and sentiment analysis outcome are evaluated by a human annotated sample, which are based on MR1 and MR2 inter agreement results. Therefore, this serves the formation of the "Gold-Standard" based on the agreed results on the sample from each dataset, which is evaluated against the majority vote (all dictionaries) category to determine its agreement level. In section 6.4.3 we will discuss the formation of the "Gold Standard".

6.4.3 Formation of Gold Standard

The 'Gold Standard' is a standard that is accepted to be a reliable and accurate reference to measure those qualities in other datasets and conclusions will be drawn about the optimal sentiment model. The Gold Standard is made up of MR1 and MR2 agreed results, which will be measured against the majority voting (all dictionaries) category to understand the strength of their agreement. The evaluation will include the analysis of precision, recall, F1 score and proportion that agreed between each sentiment category.

6.4.4 Precision and Recall Results

In Table 7 the F1 scores based on the precision and recall results have generally performed well, from 2015 MMM 0.72, and shows a steady decline with the remaining datasets, where the lowest is Anti-Austerity on 0.61. In the investigation of sentiment breakdown, most datasets have a higher negative and neutral F1 score except for Dover which continuously has a negative outlook. Furthermore, Anti-Austerity has a lower score as it has a lower negative F1 score when compared to the other datasets. Both MMM results have resulted in the same Macro F1 score of 0.66, but the difference is in the Micro F1 score with 2015 MMM on 0.72 and 2016 MMM on 0.67. Additionally, Dover has performed less well as expected due to its higher imbalance with a focus on negative, but despite Anti-Austerity having the lowest F1 score from a general perspective, it performs better in the macro/micro F1 results, as the numbers are in the early 0.60s compared with Dover Micro on 0.54 and Macro on 0.34.

In Table 187 (refer to appendix 10.14), Gold Standard and Majority Voting are 71.67% in agreement with the remaining 28.33% in disagreement. This shows both are largely in agreement, which is further evidenced by the moderate agreement shown with Cohen Kappa's 'unweighted' of 0.50, 'squared' of 0.45 and 'equal' of 0.50 with a mean result of 0.48. Additionally, Krippendorff's Alpha result of 0.50 shows a moderate level of reliability, which supports Cohen Kappa's outcome. The Z value for all three Cohen Kappa's results show a range of positive 14 to 20.4 standard deviations from the mean. The distribution is not normal and has a positive skew. Additionally, the p-value for

Cohen Kappa's is 0, which means the results are statistically significant, thus the appraiser agreement is significantly varied from what could be achieved by chance.

Precision and recall results for MR1 & MR2 compared with Majority Voting				
Overview	**MMM2015**	**MMM2016**	**Dover2016**	**Anti-Austerity2016**
Precision	0.76	0.69	0.71	0.61
Recall	0.68	0.64	0.62	0.61
F-measure	**0.72**	**0.67**	**0.66**	**0.61**
Senitment Breakdown				
Negative Precision	0.83	0.70	0.85	0.59
Negative Recall	0.73	0.62	0.60	0.58
Negative F-Measure	**0.78**	**0.66**	**0.70**	**0.59**
Neutral Precision	0.70	0.66	0.25	0.58
Neutral Recall	0.85	0.84	0.21	0.91
Neutral F-Measure	**0.77**	**0.74**	**0.23**	**0.71**
Positive Precision	0.74	0.88	0.25	0.93
Positive Recall	0.20	0.30	0.02	0.24
Positive F-Measure	**0.31**	**0.44**	**0.03**	**0.38**
Macro/ Micro Category				
Micro-Precision	0.76	0.69	0.71	0.61
Micro-Recall	0.68	0.64	0.44	0.61
Micro-F-Measure	**0.72**	**0.67**	**0.54**	**0.61**
Macro-Precision	0.76	0.75	0.45	0.70
Macro-Recall	0.59	0.59	0.28	0.58
Macro-F-Measure	**0.66**	**0.66**	**0.34**	**0.63**

Table 7 Gold Standard vs Majority Voting Precision Recall Results

In Table 188 (refer to appendix 10.14), Gold Standard and Majority Voting are 66.79% in agreement with the remaining 33.21% in disagreement. This shows both are largely in agreement, which is further evidenced by the moderate agreement shown with Cohen Kappa's 'unweighted' of 0.46, 'squared' of 0.43 and 'equal' of 0.45 with a mean result of 0.45. Additionally, Krippendorff's Alpha result of 0.44 shows a moderate level of reliability, which supports Cohen Kappa's outcome. The Cohen Kappa's mean result is less than 2015 MMM by 0.03 and Krippendorff's Alpha by 0.05, which further supports less agreement in the data. The Z value for all three Cohen Kappa's results show a range of positive 13.4 to 18.7 standard deviations from the mean. The distribution is not normal and has a positive skew, but this is less when compared to 2015 MMM. Additionally, the p-value for Cohen Kappa's is 0, which is exactly the same outcome as 2015 MMM, thus the results are statistically significant as the appraiser agreement is significantly varied from what could be achieved by chance.

6.4.5 Gold Standard Inter Rater Agreement Results

In Table 189 (refer to appendix 10.14), Gold Standard and Majority Voting are 53.93% in agreement with the remaining 46.07% in disagreement. This shows both are near equal agreement/disagreement, but agreement has the edge with the majority just over 50%. This is further evidenced by the 'slight agreement' shown with Cohen Kappa's 'unweighted' of 0.045, 'squared' of 0.029 and 'equal' of 0.037 with a mean result of 0.037. This weak outcome is supported by Krippendorff's Alpha result of -0.038 which classifies 'no agreement', thus a poor level of reliability. The Cohen Kappa's mean result is significantly lower than both MMM events, which further supports less agreement in the data. The Z value for all three Cohen Kappa's results show a range of positive 1.27 to 2.1 standard deviations from the mean. The distribution is not normal and has a slight positive skew, which is considerably less compared to both MMM events. Additionally, the p-value for Cohen Kappa's mean value is 0.10, which is not statistically significant and indicates there is not enough evidence that the appraiser agreement is different from what could be achieved by chance. These statistics show a strong outcome that supports a wide disagreement between 'Gold Standard' and 'Majority Voting'.

In Table 190 (refer to appendix 10.14), Gold Standard and Majority Voting are 61.17% in agreement with the remaining 38.93% in disagreement. This shows both are nearly two-thirds in agreement. This is further evidenced by the 'fair agreement' (which is close to 'moderate agreement' of 0.4-0.6) shown with Cohen Kappa's 'unweighted' of 0.37, 'squared' of 0.37 and 'equal' of 0.37 with a mean result of 0.37. This is supported by Krippendorff's Alpha result of 0.35 showing a fair level of reliability. The Cohen Kappa's mean result is less than 2015 MMM by 0.11 and 2016 MMM by 0.08, which is further supported by Krippendorff's Alpha by 0.15 and 0.09. This further supports less agreement in the data compared to both MMM events. The Z value for all three Cohen Kappa's results show a range of positive 12.3 to 17.1 standard deviations from the mean. The distribution is not normal and has a positive skew, which is similar to both MMM events. Additionally, the p-value for Cohen Kappa's is 0, which is exactly the same outcome as both MMM events, thus the results are statistically significant as the appraiser agreement is significantly varied from what could be achieved by chance.

The results from 'Gold Standard' shows a high level of agreement for 3 of the datasets except for Dover where it is near even on the agreement/disagreement. This comparison has shown that 'Gold Standard' can be used as a baseline against new data in the hybrid approach. The next step in this process is to implement the hybrid approach and evaluate the outcome.

6.5 Hybrid Approach

The first step of the hybrid approach is the dictionary phase as outlined in Figure 3.3 in section 3.1.6.1. To retrieve the results of precision and recall, we had to measure all the lexicon dictionaries against the manually classified category to determine the accuracy of each lexicon results. Each lexicon dictionary is categorised into True Positive (TP), True Negative (TN), False Negative (FN) and False Positive (FP) for each sentiment category for negative, positive and neutral.

In section 6.5.1 the first part focuses on 'Combined Dictionary' (refer to section 5.9 for combined dictionary setup) as there was an immediate problem identified with its results when compared with the other dictionary results. This was resolved with a series of methods to improve the reliability of the results. The proceeding second part in section 6.5.2 will explore precision, recall and f-measure, and will break this down for each sentiment category and examine the macro/micro precision, recall and f-measure (which are defined in section 5.10) for both MR1 and MR2.

6.5.1 Combined Dictionary Problem

Some of the initial sentiment analysis ran on the tweets showed promising results for the dictionaries as in 0.60s, except the combined dictionary presented a problem as there is no sentiment category classified as neutral because the scores are near 0. The tweets near score of 0 on closer inspection show many tweets should be classified as neutral. The team decided to implement a cut-off threshold, but where to cut off had to be determined. A series of different thresholds were created ranging from 1 to -1 to identify a more evenly balanced sentiment classification. Each cut-off point was run through a confusion matrix to determine the precision, recall and accuracy of each one's result. The one with the more evenly balanced precision and recall for each dataset will be chosen as the cut-off point. As a result, there are tweets that belong to the neutral category. These results would be included into a table to compare with the manually classified sentiment.

In Table 8, the original results with no threshold tended to be lowest accuracy except for Dover being its highest accuracy, which is due to the data mainly being negative. Additionally, the no threshold results for neutral are of 0 recall and precision 1 as no neutral results exists. The highest accuracy for all datasets tended to be 1, but the unevenness between the precision and recall is high. The higher accuracy reduces precision for negative and recall increases, putting this out of balance across the sentiment categories. Thus, a lower accuracy 0.5 cut-off, produces the best performance with a more evenly spread precision and recall across the sentiment categories for each dataset. The cut-off range of 0.5 to -0.5 is chosen for the combined dictionary to classify tweets as neutral.

2015 MMM 0.0	Negative	Neutral	Positive	Overall	2016 Dover 0.0	Negative2	Neutral3	Positive4	Overall5
Precision	0.71	NA	0.09	0.80	Precision	0.89	NA	0.05	0.94
Recall	0.72	0.00	0.72	1.44	Recall	0.70	0.00	0.90	1.59
F1	0.71	NA	0.17	0.88	F1	0.78	NA	0.09	0.88
Balanced Accuracy	0.67	0.50	0.65	1.82	Balanced Accuracy	0.66	0.50	0.77	1.93
2015 MMM 0.5					**2016 Dover 0.3**				
Precision	0.77	0.52	0.13	1.43	Precision	0.91	0.27	0.06	1.24
Recall	0.59	0.45	0.57	1.60	Recall	0.63	0.26	0.90	1.79
F1	0.67	0.48	0.21	1.36	F1	0.74	0.27	0.12	1.12
Balanced Accuracy	0.68	0.60	0.67	1.95	Balanced Accuracy	0.67	0.56	0.81	2.05
2015 MMM 1.0					**2016 Dover 0.5**				
Precision	0.85	0.51	0.20	1.56	Precision	0.92	0.25	0.07	1.24
Recall	0.41	0.80	0.43	1.64	Recall	0.57	0.41	0.86	1.84
F1	0.56	0.62	0.28	1.45	F1	0.70	0.31	0.13	1.14
Balanced Accuracy	0.66	0.65	0.66	1.98	Balanced Accuracy	0.67	0.59	0.82	2.08
2016 MMM 0.0	**Negative**	**Neutral**	**Positive**	**Overall**	**2016 Anti-Austerity 0.0**	**Negative**	**Neutral**	**Positive**	**Overall**
Precision	0.61	1.00	0.14	1.75	Precision	0.63	NA	0.16	0.78
Recall	0.68	0.00	0.80	1.49	Recall	0.59	0.00	0.93	1.52
F1	0.64	0.00	0.24	0.88	F1	0.61	NA	0.27	0.88
Balanced Accuracy	0.66	0.50	0.67	1.83	Balanced Accuracy	0.64	0.50	0.71	1.84
2016 MMM 0.5					**2016 Anti-Austerity 0.5**				
Precision	0.69	0.55	0.18	1.42	Precision	0.73	0.57	0.19	1.49
Recall	0.55	0.40	0.62	1.58	Recall	0.45	0.40	0.83	1.69
F1	0.61	0.47	0.27	1.35	F1	0.55	0.47	0.31	1.34
Balanced Accuracy	0.67	0.56	0.68	1.91	Balanced Accuracy	0.65	0.59	0.74	1.97
2016 MMM 1.0					**2016 Anti-Austerity 1.0**				
Precision	0.76	0.54	0.24	1.54	Precision	0.78	0.50	0.22	1.50
Recall	0.37	0.73	0.46	1.56	Recall	0.29	0.62	0.69	1.60
F1	0.50	0.62	0.32	1.44	F1	0.42	0.55	0.34	1.31
Balanced Accuracy	0.64	0.60	0.66	1.90	Balanced Accuracy	0.61	0.57	0.72	1.90

Table 8 Combined dictionary precision and recall results to set threshold (Manual)

In Table 9 the results from the combined dictionary contains no neutrals similar to the manually coded data as shown in Table 8. The same process for the cut-off was repeated for the automatically selected relevant data and 0.5 seemed an unreasonable choice, so further investigation is required to find suitable balance between precision, recall and f-measure. Table 9 shows there was an incline in accuracy within both 0.5 and 1, but the highest appears to be between the ranges of 0.6 to 1. The highest accuracy for both 2016 MMM and 2016 Anti-Austerity is 1, but except for the 2016 Dover which peaked at 0.3 and 2015 MMM at 0.9. The spread of precision, recall and f-measure is reasonably balanced at 1, so it was decided to use a cut-off of 1. However, to be consistent with the manual coded cut-off decided was used as a cut-off of 0.5 for the dataset because manual coded relevant data was more unevenly balanced in the sentiment scores for each category as set any higher compared to automated coded relevant data outcome.

[Intentionally Left Blank]

2015 MMM 0.1	Negative	Neutral	Positive	Overall	Dover 0.1	Negative2	Neutral3	Positive4	Overall5
Precision	0.70	0.11	0.64	1.454	Precision	0.80	0.07	0.62	1.49
Recall	0.47	0.48	0.43	1.383	Recall	0.56	0.40	0.45	1.41
F1	0.56	0.17	0.49	1.211	F1	0.65	0.11	0.49	1.25
Balanced Accuracy	0.64	0.55	0.62	1.816	Balanced Accuracy	0.67	0.54	0.65	1.86
2015 MMM 0.5					**Dover 0.5**				
Precision	0.59	0.38	0.52	1.491	Precision	0.71	0.29	0.52	1.52
Recall	0.52	0.45	0.48	1.449	Recall	0.59	0.40	0.48	1.47
F1	0.54	0.39	0.47	1.400	F1	0.63	0.31	0.47	1.42
Balanced Accuracy	0.65	0.55	0.63	1.839	Balanced Accuracy	0.67	0.55	0.65	1.87
2015 MMM 0.9					**Dover 0.9**				
Precision	0.46	0.65	0.38	1.492	Precision	0.54	0.61	0.36	1.51
Recall	0.57	0.45	0.55	1.568	Recall	0.65	0.39	0.54	1.58
F1	0.50	0.50	0.42	1.425	F1	0.58	0.44	0.41	1.43
Balanced Accuracy	0.67	0.58	0.65	1.901	Balanced Accuracy	0.68	0.57	0.67	1.91
2015 MMM 1.0					**Dover 1.0**				
Precision	0.43	0.69	0.35	1.471	Precision	0.51	0.66	0.34	1.51
Recall	0.59	0.44	0.56	1.585	Recall	0.67	0.39	0.55	1.61
F1	0.49	0.51	0.41	1.401	F1	0.57	0.46	0.40	1.42
Balanced Accuracy	0.67	0.58	0.66	1.907	Balanced Accuracy	0.68	0.57	0.67	1.92
2016 MMM 0.1	**Negative**	**Neutral**	**Positive**	**Overall**	**Anti-Austerity 0.1**	**Negative**	**Neutral**	**Positive**	**Overall**
Precision	0.71	0.08	0.68	1.47	Precision	0.63	0.07	0.82	1.52
Recall	0.46	0.46	0.42	1.35	Recall	0.44	0.47	0.51	1.42
F1	0.55	0.12	0.49	1.17	F1	0.51	0.11	0.60	1.23
Balanced Accuracy	0.65	0.53	0.62	1.80	Balanced Accuracy	0.66	0.55	0.65	1.86
2016 MMM 0.5					**Anti-Austerity 0.5**				
Precision	0.58	0.34	0.57	1.49	Precision	0.50	0.32	0.72	1.55
Recall	0.51	0.46	0.46	1.43	Recall	0.51	0.46	0.54	1.51
F1	0.53	0.37	0.48	1.38	F1	0.49	0.36	0.59	1.44
Balanced Accuracy	0.66	0.54	0.63	1.83	Balanced Accuracy	0.68	0.56	0.65	1.89
2016 MMM 0.9					**Anti-Austerity 0.9**				
Precision	0.44	0.66	0.41	1.51	Precision	0.37	0.53	0.58	1.48
Recall	0.58	0.47	0.53	1.58	Recall	0.57	0.44	0.57	1.57
F1	0.49	0.52	0.43	1.45	F1	0.44	0.45	0.55	1.44
Balanced Accuracy	0.68	0.58	0.65	1.91	Balanced Accuracy	0.70	0.56	0.64	1.89
2016 MMM 1.0					**Anti-Austerity 1**				
Precision	0.42	0.69	0.38	1.48	Precision	0.34	0.61	0.53	1.48
Recall	0.59	0.46	0.54	1.59	Recall	0.58	0.43	0.57	1.58
F1	0.48	0.53	0.42	1.42	F1	0.42	0.47	0.52	1.41
Balanced Accuracy	0.68	0.58	0.65	1.91	Balanced Accuracy	0.70	0.56	0.63	1.89

Table 9 Combined dictionary precision and recall results to set threshold (automated)

6.5.1.1 Majority vote for all dictionaries

To determine the majority vote for all dictionaries, the scores for each dictionary are classed as either positive, negative or neutral sentiment for each tweet. This will enable a total count for each sentiment category to identify the majority sentiment. If there is no majority consensus, then a program will decide with flip of a coin with the categories that are tied to determine the majority.

6.5.1.1.1 Manually Coded

The manually coded data results for the majority vote are outlined in Table 10. A description of the results are as follows: -

- 2015 MMM with the highest vote is for negative, with neutral not far behind and the least is positive.
- 2016 MMM positive, neutral and negative are close within each other, but the highest vote is neutral, and the least being negative.
- 2016 Dover had a clear majority of negative, which is followed by approximately half the vote each for positive and neutral, with least being neutral.
- 2016 Anti-Austerity positive, negative and neutral are closely aligned, but the majority is neutral, followed by positive and negative.

The most divisive demonstration is 2016 Dover due to the topic based on immigration, which was hotly debated in politics in that period of time. In contrast with the other cases, MMM (about capitalism) and Anti-Austerity are broader in topic and that gathers wider public support. These may be the reasons why there are differences in the sentiment.

	(Total Vote)	**(Total Vote)**
Sentiment Category	**MMM 2015**	**MMM 2016**
Positive	747 (22.66%)	1065 (31.73%)
Negative	1439 (43.66%)	1044 (31.11%)
Neutral	1110 (33.58%)	1247 (37.16%)
Total	3296	3356
	Dover 2016	**Anti-Austerity 2016**
Positive	729 (25.76%)	1819 (33.40%)
Negative	1425 (50.25%)	1644 (30.19%)
Neutral	676 (23.89%)	1983 (36.41%)
Total	2830	5446

Table 10 Results of the majority voting for each dataset

The next phase is to determine whether any of the dictionaries required exclusion depending on the quality of its precision, recall, F1 and its total vote.

6.5.1.1.2 Automated Coding

The results in Table 11 show the proportion of each sentiment for the automatically coded data, automatically coded for both relevant/irrelevant and sentiment (using the majority vote of the dictionaries): -

Majority Voting Results		
	(Total Vote)	**(Total Vote)**
Sentiment Category	**MMM 2015**	**MMM 2016**
Positive	8133 (27.65%)	4523 (29.20%)
Negative	10277 (34.93%)	5043 (32.55%)
Neutral	11010 (37.42%)	5925 (38.25%)
Total	29420	15491
	Dover 2016	**Anti-Austerity 2016**
Positive	748 (23.57%)	12854 (42.90%)
Negative	1605 (50.57%)	6506 (21.71%)
Neutral	821 (25.87%)	10603 (35.39%)
Total	3174	29963

Table 11 Majority voting with 0.5 cut off (automated)

Overall, three datasets' results' show negative and neutral have the majority proportion for the tweets classified, but with the exception where Anti-Austerity has the highest level of positive sentiment. The manual coded relevant data compared to automated are similar to Dover, AA positive and neutral higher than negative, and both MMM events are similar except 2015 MMM negative is higher with positive and neutral lower. The next phase is to determine whether any of the dictionaries required exclusion depending on the quality of its precision, recall, F1 and its total vote.

6.5.1.1.3 Exclusion of dictionaries

The exclusion of dictionaries is to improve the machine learning classification. The poorest dictionaries results were chosen based on their precision, recall and f-measure results produced based against the MR1. The dictionaries identified for possible removal are as follows: -

- Slangsd
- Senticent
- Socal Google
- Berkeley

In Table 12, manually coded relevant data shows the removal of the specific dictionaries have little effect on the outcome when compared with Table 10 results as these dictionaries were somewhat in agreement. Therefore, the manual coded (relevant) data's majority vote results show to potentially keep the dictionaries in with the results. The next step is to verify whether the automated process shows little effect to exclude these dictionaries.

Sentiment Category	MMM 2015	MMM 2016
Positive	687 (20.84%)	1005 (29.95%)
Negative	1463 (44.39%)	1049 (31.26%)
Neutral	1145 (34.74%)	1302 (38.80%)
Total	3296	3356
	Dover 2016	**Anti-Austerity 2016**
Positive	689 (24.35%)	1739 (31.93%)
Negative	1428 (50.46%)	1647 (30.24%)
Neutral	713 (25.19%)	2060 (37.83%)
Total	2830	5446

Table 12 Total vote for each sentiment category – removal of dictionaries (manual)

The automated coded (relevant) data displayed in Table 13, shows all datasets increased in negativity (except for AA which decreased), a decrease for neutral (increase slightly for AA), and in both 2015 MMM and AA saw an increase in positivity in the categories compared with Table 12.

Sentiment Category	MMM 2015	MMM 2016
Positive	7706 (22.19%)	4174 (26.94%)
Negative	10275 (34.93%)	5129 (33.11%)
Neutral	11439 (38.88%)	6188 (39.95%)
Total	29420	15491
	Dover 2016	**Anti-Austerity 2016**
Positive	694 (21.87%)	12147 (40.54%)
Negative	1620 (51.04%)	6512 (21.73%)
Neutral	860 (27.10%)	11304 (37.73%)
Total	3174	29963

Table 13 Total vote for each sentiment category – removal of dictionaries (automated)

Table 13 results for the automated way has similarly shown little change when compared with Table 11, as these dictionaries were somewhat in agreement. Therefore, it has been decided to keep these dictionaries to be used in the machine learning phase. The next step is to explore the initial sentiment analysis results based on the lexicon approach.

6.5.2 Dictionary: Breakdown of Precision and Recall Results

Both MR1 and MR2 dictionary results provides a breakdown of precision (in section 6.5.2), recall and f-measure based on each sentiment category. Afterwards, in section 6.5.3 both MR1 and MR2 is validated by use macro/micro precision, recall and f-measure. The numbers presented in these results are rounded up or down based on the figures in each of the tables. These results will help to understand the strength of the dictionaries results to identify which is the best and worst performing ones.

Both MR1 and MR2 precision, recall and f-measure (refer to definitions in section 5.10) will be explored for each dataset for each sentiment category.

6.5.2.1 MR1 Results

6.5.2.1.1 Negative

Table 14 shows the range of F1 scores for negative sentiment using each of the dictionaries, whilst Table 15 again summarises which dictionaries performed best and worst.

F-measure Range					
	0.80s	0.70s	0.60s	0.50s	Below 0.50s
2015 MMM	2	3	8	1	5
2016 MMM	0	0	6	6	6
2016 Dover	0	6	5	6	1
2016 Anti-Austerity	0	0	8	7	4

Table 14 Negative F-measure range between each dataset for MR1 for each of the 19 dictionaries

In Table 15 it can be seen that the dictionaries struggle more with AA and 2016 MMM where the F1 scores are generally lower. In Table 15 it can be seen that "Jockers Rinkers" is the highest or second highest in terms of F1 in all cases, although this seems to be because of a high recall in all except for AA. The worst performing dictionaries consistently is "Socal Google" for F1, but for all datasets precision has a different dictionary (both AA and 2016 MMM is Slangsd", 2015 MMM is "StentiStrength" and Dover is "Stanford") only for its category for lowest, which most sit around 0.50 score and similar score in recall, which is a poor result.

MMM 2015	**1**	**2**	**3**	**Lowest**
F-measure	Jockers Rinker and Syuzhet Jockers 0.80	Sentimentr Jockers 0.79	Vadar 0.71.	Socal Google 0.31
Precision	Syuzhet NRC and Inquirer 0.87	Bing and Huliu 0.86	Sentiment NRC 0.84	SentiStrength 0.56
Recall	Jockers Rinker 0.78	Syuzhet Jockers and Sentimentr Jockers on 0.77	Vadar 0.62	Socal Google 0.20
MMM 2016				
F-measure	Syuzhet Jockers 0.70	Jockers Rinker and Sentimentr Jockers 0.69	Vadar 0.66	Socal Google 0.31
Precision	Vadar 0.80	Bing 0.78	Afinn 0.77	Slangsd 0.49
Recall	Syuzhet Jockers and Jockers Rinker 0.67	Sentimentr Jockers 0.65	Vadar 0.57	Socal Google 0.18
Dover 2016				

F-measure	Jockers Rinker 0.80	Syuzhet Jockers and Sentimentr Jockers 0.79	Sentistrength 0.78	Socal Google 0.39
Precision	Syuzhet NRC and Sentiment NRC 0.97	Bing, Huliu and SentiStrength 0.96	Sentimentr Jockers 0.95	Stanford 0.85
Recall	Jockers Rinker 0.69	Syuzhet Jockers and Sentimentr Jockers 0.68	SentiStrength 0.66	Socal Google 0.24
Anti-Austerity 2016				
F-measure	Syuzhet Jockers, Sentimentr Stanford, SentiStrength and Sentimentr Jockers 0.67	Jockers Rinker 0.66	Vadar 0.58	Socal Google 0.27
Precision	Bing 0.84	Huliu 0.83	Syuzhet NRC, Sentiment NRC and Vadar on 0.82	Slangsd 0.52
Recall	Stanford 0.61	Jockers family and SentiStrength 0.57	Slangsd 0.47	Socal Google 0.16

Table 15 Negative precision/recall summary for best to worst dictionaries for MR1

MR1 results show that one or more of the Jockers family performs best in most cases throughout for F1 and differs greatly in the third position, which suggests Vadar performs best except for Dover with "SentiStrength". This shows some dictionaries are stronger performers when it comes specifically towards negative sentiment. However, there is agreement between negative breakdown that both "Vadar" and "SentiStrength" appears in the top 3 positions.

6.5.2.1.2 Neutral

Table 14 shows the range of F1 scores for neutral sentiment using each of the dictionaries, whilst Table 15 again summarises which dictionaries performed best and worst.

F-measure Range					
	0.80s	0.70s	0.60s	0.50s	Below 0.50s
2015 MMM	0	0	9	3	7
2016 MMM	0	0	10	5	4
2016 Dover	0	0	0	0	19
2016 Anti-Austerity	0	0	8	7	4

Table 16 Neutral F-measure range between each dataset for MR1

In Table 156 it can be seen that the dictionaries struggle more with Dover and AA where the F1 scores are generally lower with lowest being Dover with all 19 dictionaries scoring below

0.50. In Table 157 it can be seen that “SentiStrength” is the highest in terms of F1 in all cases except for 2015 MMM is “Vadar”, although this seems to be because of a high precision for in all except for 2015 MMM. The “SentiStrength” dictionary has lower recall of 0.34 for Dover, but for both 2016 MMM and AA recall is about 0.10 lower than their precision scores. These low recall scores with a higher precision suggests “SentiStrength” performs well on Dover dataset. Whereas both 2016 MMM and AA which indicates similar number of false positives as the number of false negatives are near equally important this provides a weak result. Furthermore, Vadar” suffers a similar issue, but is different as recall is 0.70 higher than precision on 0.67. The worst performing dictionaries consistently are “Berkeley” and “Senticnet” for F1, and is similar for both precision and recall for all datasets except for Dover recall with the inclusion of “Stanford”, which most for precision is below 0.09 and recall below 0.33 score except for Dover on 0.17 and similar score in recall, which is a very poor result.

MMM 2015	**1**	**2**	**3**	**Lowest**
F-measure	Vadar 0.68	Huliu and Bing 0.67	Afinn 0.66	Berkeley 0.067 and Senticnet 0.089
Precision	Loughran Mcdonald 0.82	Bing and Huliu 0.79	Inquirer 0.78	Berkeley 0.038 and Senticnet 0.048
Recall	Jockers Rinker 0.84	Syuzhet Jockers 0.83	Sentimentr Jockers 0.82	Berkeley 0.29 and Senticnet is 12th position
MMM 2016				
F-measure	Sentistrength 0.69	Loughran Mcdonald, Afinn, Vadar and Huliu is 0.68	Bing 0.67	Senticnet 0.12 and Berkeley 0.13
Precision	Loughran Mcdonald 0.81	Inquirer 0.75	SentiStrength 0.74	Senticnet 0.063 and Berkeley 0.084
Recall	Jockers Rinker 0.83	Sentimentr Jockers and Syuzhet Jockers 0.81	Vadar 0.75	Berkeley 0.33 and Senticnet is in 5th position
Dover 2016				
F-measure	SentiStrength 0.44	Vadar 0.43	Sentimentr Jockers and Syuzhet Jockers 0.41	Berkeley 0.10 and Senticnet 0.11
Precision	Loughran Mcdonald 0.80	Inquirer 0.66	SentiStrength and Sentiment NRC 0.62	Senticnet 0.067 and Berkeley 0.075
Recall	Syuzhet Jockers and Sentimentr Jockers 0.46	Jockers Rinker 0.45	Vadar 0.40	Berkeley and Stanford 0.17 and Senticnet is 5th position.
Anti-Austerity 2016				

F-measure	SentiStrength 0.69	Huliu and Loughran Mcdonald 0.66	Bing and Inquirer 0.65	Berkeley 0.06 and Senticnet 0.10
Precision	Loughran Mcdonald 0.86	Sentistrength 0.74	Bing and Inquirer 0.71	Berkeley 0.03 and Senticnet 0.05
Recall	Senticnet and Jockers Rinker 0.74	Sentimentr Jockers and Syuzhet Jockers 0.73	Vadar 0.68	Berkeley 0.31 and Senticnet is 1st position

Table 17 Neutral precision/ recall summary for best to worst dictionaries for MR1

In 2016 Dover F1 ranges from 0.44 to 0.10, with the highest being SentiStrength 0.44 and lowest Berkeley 0.10 with 0.34 difference, showing a moderate gap between their positions. Both MMM F1s are far higher compared with Dover, and 2016 MMM and Dover agree that SentiStrength is the highest dictionary. Both 2015 MMM and Dover agree the lowest is Berkeley. The precision ranges from 0.80 to 0.07, with the highest being Loughran McDonald 0.80 and the lowest Senticnet 0.07 with 0.73 difference, showing a very large gap between their positions. Precision for Dover is similar to both MMM score ranges, and Dover agrees with both MMM that Loughran MacDonald is the highest, but is in less agreement with the lowest as both MMM are Berkeley, whereas Dover is Senticnet. Overall Dover has a slightly smaller precision range. The recall ranges from 0.47 to 0.17, where the highest are both Syuzhet Jockers and Sentimentr Jockers on 0.46 and lowest both Berkeley and Stanford on 0.17 with 0.30 difference, showing a moderate difference between their positions. Both MMM and Dover similarly agrees that the Jockers Family are the highest recall, and the lowest is Berkeley.

In 2016 Anti-Austerity F1 ranges from 0.69 to 0.06, with the highest being SentiStrength 0.69 and lowest Berkeley 0.06 with 0.63 difference, showing a very large gap between their positions. The F1 range is similar to both MMM results except for Dover which is the lowest. However, sentiment dictionaries responded best in the negative category which means most tweets are classified negative, therefore, the numbers for neutral were always going to be lowered, as there are less tweets deemed neutral. The precision ranges from 0.86 to 0.03, with the highest being Loughran McDonald 0.86 and lowest Berkeley 0.03 with 0.83 difference, showing a very large gap between their positions. The highest precision for Anti-Austerity is the largest compared to the other datasets and all datasets are in agreement that Loughran MacDonald is ranked 1 and is in agreement with both MMM that Berkeley is the lowest, except for Dover with Senticnet. The recall ranges from 0.74 to 0.31, where the highest are both Senticnet and Jockers Rinker on 0.74 and lowest is Berkeley on 0.31 with 0.33 difference, showing a moderate difference between their positions. Anti-Austerity has the second lowest recall range, and some dictionaries for the Jocker Family are highest, but with an exception where Anti-Austerity has two that are the top, the other of which is Senticnet, which is one of the lowest performing for most instances in MR1 and MR2 results.

6.5.2.1.3 Positive

Table 18 shows the range of F1 scores for neutral sentiment using each of the dictionaries, whilst Table 19 again summarises which dictionaries performed best and worst.

F-measure Range					
	0.80s	0.70s	0.60s	0.50s	Below 0.50s
2015 MMM	0	0	0	0	19
2016 MMM	0	0	0	1	18
2016 Dover	0	0	0	0	19
2016 Anti-Austerity	0	0	0	1	18

Table 18 Positive F-measure range between each dataset for MR1

In Table 18 it can be seen that the dictionaries struggle with all datasets as the F1 scores are very low, thus poor in result. In Table 19 it can be seen that "SentiStrength" is the highest in terms of F1 in all cases except for 2015 MMM although this seems to be because of a precision is high and a low recall. The worst performing dictionaries consistently are mainly "Slangsd" and "Berkeley" for F1 (Stanford and Sentiword do appear but less often), and both "Stanford" and "Slangsd" is similarly identified for precision and recall except "Stanford" is replaced with "Socal Google"/"Sentiword"/"Berkeley" for recall alongside "Slangsd". These dictionaries tend perform below a precision of 0.37 and a recall of 0.32 with lowest for both on 0 for "Slangsd" based on Dover dataset, which is very poor.

MMM 2015	**1**	**2**	**3**	**Lowest**
F-measure	Bing and Huliu 0.38	Afinn 0.36	Vadar 0.35	Slangsd 0.07 and Stanford 0.09
Precision	Vadar 0.84	Sentimentr Jocker 0.83	Jockers Rinker, Syuzhet Jockers and Senticnet 0.81	Stanford 0.11 and Slangsd 0.13
Recall	Loughran McDonald 0.28	Bing 0.27	Huliu 0.26	Slangsd 0.13 and Socal Google 0.06
MMM 2016				
F-measure	Sentistrength 0.50	Bing and Huliu 0.43	Inquirer 0.42	Slangsd 0.074 and Berkeley 0.18
Precision	Vadar 0.87	Jockers Rinker 0.86	Sentimentr Jockers 0.85	Slangsd 0.087 and Stanford 0.21
Recall	SentiStrength 0.40	Loughran Mcdonald 0.39	Bing 0.31	Slangsd 0.065 and Berkeley 0.10
Dover 2016				
F-measure	SentiStrength 0.19	Bing and Loughran Mcdonald 0.15	Combined Dictionary and Huliu 0.13	Slangsd 0.00 and Sentiword 0.059

Precision	Bing 0.97	Syuzhet Jockers, Sentimentr Jockers and Senticnet 0.90	Combined Dicitonary, Vadar and Jockers Rinker 0.86	Slangsd 0.00 and Loughran Mcdonald 0.35
Recall	SentiStrength 0.11	Loughran Mcdonald 0.093	Bing 0.086	Slangsd 0.00 and Sentiword 0.32
Anti-Austerity 2016				
F-measure	SentiStrength 0.55	Huliu and Bing 0.42	Stanford 0.41	Slangsd 0.077 and all Socal Google, Sentiword and Senticnet 0.22
Precision	Syuzhet Jockers 0.95	Sentimentr Jockers and Jockers Rinker 0.94	Vadar and Berkeley 0.91	Slangsd 0.087 and Loughran Mcdonald 0.37
Recall	SentiStrength 0.42	Loughran Mcdonald 0.37	Stanford 0.34	Slangsd 0.069, all Senticnet, Sentiword and Berkeley 0.13

Table 19 Positive precision/recall summary for best to worst dictionaries for MR1

MR1 positive results show that "SentiStrength" performs best in most cases throughout for F1 except for 2015 MMM joint top with "Bing" and "Huliu" and differs greatly in both second and third position with either "Bing", "Afinn", "Huliu" and "Vadar" in second or third position. For the first time, the combined dictionary appeared in joint third with Huliu for F1 based on Dover dataset, however, the results are very poor. Again, MR1 for neutral agrees with negative breakdown that "Bing" and "Afinn" are highly ranked and that the top ranked dictionaries can vary as shown in all sentiment categories compared with the overall scores in section 6.5.2. Overall, the first top positioned dictionary differ from each specific sentiment category, but there is agreement that "Afinn", "Bing" and "Vadar" are consistently in the top 3 positions for each sentiment category and overall scores section 6.5.2.

[Intentionally Left Blank]

6.5.2.1.4 Summary

The strongest performance of F1 scores for negative are one or more Jockers family, but AA includes “Stanford” and “SentiStrength” as well. For both neutral and positive categories “SentiStrength” is consistently has the best performance except for 2015 MMM where both neutral (Vadar) and positive (Bing/Huliu) is strongest.

Top f-measure for each dataset sentiment category			
Dataset	**Negative**	**Neutral**	**Positive**
2015 MMM	Jockers Rinker and Syuzhet Jockers 0.80	Vadar 0.68	Bing and Huliu 0.38
2016 MMM	Syuzhet Jockers 0.70	Sentistrength 0.69	SentiStrength 0.50
2016 Dover	Jockers Rinker 0.80	SentiStrength 0.44	SentiStrength 0.19
2016 Anti-Austerity	Syuzhet Jockers, Sentimentr Stanford, SentiStrength and Sentimentr Jockers 0.67	SentiStrength 0.69	SentiStrength 0.55

Table 20 Top F-measure for each sentiment category for MR1

The worst performing dictionaries are “Combined”, “Berkeley”, “Senticnet” and “NRC” have scored the lowest below 0.3. The reason these dictionaries may have the lowest scores could be due to less words are identified by those dictionaries in each of the tweets. However, the “Combined Dictionary” has the largest set of terms but performs not that well compared to the smaller lexicons, which again may be due to some words are not scored in the sentiment outcome and/ or how balanced the scores are in the term selection as might be in favour of one or more sentiment categories that can impact the overall F1 score.

6.5.2.2 MR2 Results

6.5.2.2.1 Negative

A similar process was carried out for MR2 and Table 21 shows the range of F1 scores for negative sentiment using each of the dictionaries, whilst Table 22 again summarises which dictionaries performed best and worst.

F-measure Range					
	0.80s	**0.70s**	**0.60s**	**0.50s**	**Below 0.50s**
2015 MMM	0	0	5	8	6
2016 MMM	0	0	0	6	13
2016 Dover	0	6	4	8	1
2016 Anti-Austerity	0	0	0	0	19

Table 21 Negative F-measure range between each dataset for MR2

In Table 21 it can be seen that most dictionaries struggle with all datasets as the F1 scores are very low, but 5 dictionaries in 0.60s for 2015 MMM and 4 for Dover with 6 in the 0.70s range. In Table 22 it can be seen that one or more of the Jockers family has the highest in terms of F1 in all cases although this seems to be because of a high recall above precision. The worst performing dictionaries are both "Slangsd" and "Syuzhet NRC" for F1 except for Dover which are both "Socal Google" and "Loughran MacDonald". However, the precision and recall are not similar to F1 dictionaries results unlike MR1 negative results. The precision lowest is consistently "Slangsd" (typically under 0.38 except for Dover on 0.67), but recall lowest is "Socal Google" (mostly below 0.30), which both show a poor result.

MMM 2015	**1**	**2**	**3**	**Lowest**
F-measure	Jockers family 0.65	Afinn 0.62	Vadar 0.61	Syuzhet NRC 0.34 Socal Google 0.30
Precision	Bing, Inquirer and Huliu 0.58	Afinn and Vadar 0.57	Syuzhet Jockers and Sentimentr Jockers 0.55	SentiStrength 0.38 Slangsd 0.37
Recall	Jockers Rinker and Syuzhet Jockers 0.81	Sentimentr Jockers 0.80	Afinn 0.67	Syuzhet NRC 0.25 Socal Google 0.21
MMM 2016				
F-measure	Jockers Rinker and Sentimentr Jockers 0.56	Syuzhet Jockers and Vadar 0.55	Afinn 0.54	Syuzhet NRC 0.32
Precision	Vadar 0.50	Afinn 0.49	Bing 0.47	Stanford 0.32 Slangsd 0.30
Recall	Jockers Rinker 0.71	Syuzhet Jockers 0.70	Sentimentr Jockers 0.69	Syuzhet Jockers 0.26 Socal Google 0.16
Dover 2016				
F-measure	Jockers Rinker 0.76	Sentimentr Jockers 0.75	Syuzhet Jockers and SentiStrength 0.74	Loughran MacDonald 0.50 Socal Google 0.38
Precision	Both NRC 0.80	Bing, Inquirer, Vadar, Jockers Rinker, SentiStrength, Sentimentr Jockers and Huliu 0.79	Afinn and Syuzhet Jockers 0.78	Slangsd and Stanford 0.67
Recall	Jockers Rinker 0.72	Syuzhet Jockers and Sentimentr Jockers 0.71	SentiStrength 0.69	Loughran MacDonald 0.38 Socal Google 0.25
Anti-Austerity 2016				
F-measure	Syuzhet Jockers 0.35	SentiStrength, Combined Dictionary, Bing 0.34	Sentimentr Jockers, Jockers Rinker, Stanford, Huliu 0.33	Socal Google 0.15 Slangsd 0.21
Precision	Bing and Syuzhet NRC 0.26	Sentimentr NRC and Huliu 0.25	Combined Dictionary, Syuzhet Jockers,	Slangsd 0.14 Sentiword 0.16

			Vadar, SentiStrength 0.24	
Recall	Stanford 0.65	Syuzhet Jockers 0.61	SentiStrength and Jockers Rinker 0.59	Socal Google 0.14 Loughran MacDonald 0.30

Table 22 Negative precision/recall summary for best to worst dictionaries for MR2

MR2 negative results is similar to MR1 where it shows one or more of the Jockers family is the highest scores F1 and differs greatly in both second and third position where in AA has more dictionary occurrences than the other datasets. Similar to MR1 negative, Jockers family dominates in specifically for negative for MR2. This reiterates that some dictionaries are stronger performers when it comes specifically towards negative sentiment. However, there is agreement between negative breakdown to overall sentiment F1 that both "Vadar" and "SentiStrength" appears in the higher positions.

6.5.2.2.2 Neutral

Table 23 shows the range of F1 scores for negative sentiment using each of the dictionaries, whilst Table 24 again summarises which dictionaries performed best and worst.

F-measure Range					
	0.80s	0.70s	0.60s	0.50s	Below 0.50s
2015 MMM	0	0	7	8	4
2016 MMM	0	2	6	7	4
2016 Dover	0	0	0	0	19
2016 Anti-Austerity	0	1	8	3	7

Table 23 Neutral F-measure range between each dataset for MR2

In Table 23 it can be seen that all dictionaries perform badly for Dover, but for the other datasets performs well with nearly half of the dictionaries scores are 0.60s to 0.70s. In Table 22 it can be seen that "Loughran MacDonald" has the highest score for F1 in most cases except for 2015 MMM joint top with "Bing" and "Huliu". This seems to be because of either high recall or high precision, and both 2016 MMM and AA have nearly similar scores, which is poor. The worst performing dictionaries are both "Berkeley" and "Senticnet" for F1 and precision and recall are similar as well. The precision lowest is consistently "Senticnet" (tends to be 0.05 or under) but recall lowest is "Berkeley" (mostly below 0.50 except for AA on 0.72), which shows a poor result.

MMM 2015	**1**	**2**	**3**	**Lowest**
F-measure	Bing and Huliu 0.69	Inquirer 0.68	Loughran MacDonald 0.67	Berkeley 0.08 Senticnet 0.08
Precision	Loughran MacDonald 0.68	Bing 0.66	Inquirer and Huliu 0.65	Berkeley 0.05 Senticnet 0.04
Recall	Syuzhet Jockers and Jockers Rinker 0.90	Sentimentr Jockers 0.89	Vadar 0.81	Stanford and SentiStrength 0.58

				Berkeley 0.52
MMM 2016				
F-measure	Loughran MacDonald 0.71	Inquirer 0.70	Huliu 0.69	Berkeley 0.17 Senticnet 0.10
Precision	Loughran MacDonald 0.73	Inquirer 0.68	Bing 0.65	Berkeley 0.10 Senticnet 0.05
Recall	Jockers Rinker 0.89	Sentimentr Jockers 0.88	Syuzhet Jockers 0.87	Stanford 0.64 Berkeley 0.51
Dover 2016				
F-measure	Loughran MacDonald and Inquirer 0.48	Sentiment NRC 0.46	SentiStrength and Syuzhet NRC 0.45	Berkeley 0.11 Senticnet 0.07
Precision	Loughran MacDonald 0.69	Inquirer 0.55	Both NRC 0.51	Berkeley 0.07 Senticnet 0.04
Recall	Jockers Rinker 0.56	Sentimentr Jockers and Syuzhet Jockers 0.55	Vadar 0.53	Stanford 0.31 Berkeley 0.29
Anti-Austerity 2016				
F-measure	Loughran MacDonald 0.77	Bing 0.68	Inquirer and Huliu 0.67	Berkeley 0.08 Senticnet 0.07
Precision	Loughran MacDonald 0.73	Bing 0.57	Inquirer 0.56	Berkeley 0.04 Senticnet 0.04
Recall	Senticnet 0.93	Jockers Family 0.91	Vadar 0.88	Slangsd 0.77 Berkeley 0.72

Table 24 Neutral precision/recall summary for best to worst dictionaries for MR2

In Table 23, the 2016 Anti-Austerity f-measure rate range is 0.77 to 0.08. In Table 24, the highest is Loughran MacDonald on 0.77 and lowest is Senticnet on 0.08 with 0.69 difference, showing a large gap between their positions. The precision ranges from 0.73 to 0.04, with the highest being Loughran MacDonald on 0.73 and lowest both Berkeley and Senticnet on 0.04 with 0.69 difference, showing a large gap between their positions. The recall ranges from 0.93 to 0.72, where the highest is Senticnet on 0.93 and lowest Berkeley on 0.72 with 0.19 difference, showing a moderate gap between their positions. In the neutral category, there is a much higher recall in each dataset, where the highest is around 0.90s except for 2016 Dover. In terms of precision, this is lower precision than recall, of which the highest is mostly around 0.70s. The highest f-measure is for Loughran MacDonald, and the two lowest dictionaries are mainly both Berkeley and Senticnet. These f-measures are positioned higher than negative by approximately up to 0.10 except for Dover which is +28 with Jockers Rinker.

6.5.2.2.3 Positive

Table 25 shows the range of F1 scores for negative sentiment using each of the dictionaries, whilst Table 26 again summarises which dictionaries performed best and worst.

F-measure Range					
	0.80s	**0.70s**	**0.60s**	**0.50s**	**Below 0.50s**
2015 MMM	0	0	0	0	19
2016 MMM	0	0	0	0	19
2016 Dover	0	0	0	0	19
2016 Anti-Austerity	0	0	0	0	19

Table 25 Positive F-measure range between each dataset for MR2

In Table 25 it can be seen that all dictionaries perform badly for every dataset. In Table 26 it can be seen that "SentiStrength" and "Bing/Huliu" has the highest score for F1. This seems to be because of a high precision for Bing, and both "SentiStrength" and "Huliu" has a high precision and low recall, but the scores are poor. The worst performing dictionaries are mostly "Slangsd" and "Socal Google" for F1. The precision lowest is consistently "Slangsd" (tends to be 0.11 or under) but recall lowest is "Slangsd" (below 0.11 with most on 0.6 or lower), which shows a poor result.

MMM 2015	**1**	**2**	**3**	**Lowest**
F-measure	Bing 0.31	Huliu and Vadar 0.30	Afinn and Syuzhet Jockers 0.28	Socal Google 0.12 SentiStrength, Slangsd and Stanford 0.10
Precision	Berkeley 0.70	Senticnet 0.69	Vadar 0.65	Loughran MacDonald 0.13 Stanford 0.10
Recall	Bing and Huliu 0.22	Afinn and Vadar 0.20	Loughran MacDonald 0.19	SentiWord, SentiStrength and Berkeley 0.08 Socal Google and Slangsd 0.07
MMM 2016				
F-measure	Huliu 0.44	Bing 0.43	Vadar 0.42	Stanford 0.17 Slangsd 0.12
Precision	Senticnet 0.78	Vadar, Jockers Rinker and Sentimentr Jockers 0.72	Syuzhet Jockers 0.70	Stanford 0.16 Slangsd 0.11
Recall	SentiStrength 0.38	Loughran MacDonald 0.37	Huliu 0.35	Berkeley 0.13 Slangsd 0.12
Dover 2016				
F-measure	SentiStrength 0.20	Loughran MacDonald 0.18	Afinn 0.17	Socal Google 0.06 Slangsd 0.02

Precision	Senticnet 0.80	Sentimentr Jockers and Berkeley 0.78	Jockers Rinker 0.76	Stanford 0.27 Slangsd 0.06
Recall	Loughran MacDonald 0.13	SentiStrength 0.12	Afinn 0.10	Socal Google 0.03 Slangsd 0.01
Anti-Austerity 2016				
F-measure	SentiStrength 0.37	Afinn 0.36	Huliu and Bing 0.35	Socal Google 0.19 Slangsd 0.06
Precision	Berkeley 0.83	Senticnet 0.81	Sentiment Jockers 0.77	Loughran MacDonald 0.27 Slangsd 0.06
Recall	Loughran MacDonald, Stanford and SentiStrength 0.30	Bing, Huliu and Afinn 0.24	Inquirer 0.22	Socal Google 0.12 Slangsd 0.06

Table 26 Positive precision/recall summary for best to worst dictionaries for MR2

In the positive category, there is a much lower recall in each dataset, where the highest is around 0.30s. In terms of precision, this is higher than recall, of which the highest is mostly around 0.70s to 0.80s. Precision and recall are the opposite end of each other, which is far greater than the other sentiment categories, especially the lower recall results. The top f-measure for Dover and Anti-Austerity is SentiStrength on 0.20 and 0.37, but 2015 MMM is Bing of 0.31 and the highest f-measure is Huliu of 0.44 for 2016 MMM. Anti-Austerity's results for the two lowest dictionaries are mostly Socal Google and Slangsd, however, the results are more varied compared to the other datasets. These f-measures are the lowest compared to negative and neutral by far.

6.5.2.2.4 Summary

MR2 highest F1 scores is negative with one or more Jockers family with neutral slightly lower, and positive much lower correct classifications. However, in Table 27 neutral has the highest f-measure scores except for Dover (where negative Jockers Rinkers on 0.76 and neutral Loughran MacDonald/Inquirer 0.48) compared to both negative and positive categories.

In Table 27, the highest f-measure for negative is mostly Jockers Rinker, neutral is mainly Loughran MacDonald (except 2015 MMM with Bing and Huliu) and lastly positive are both SentiStrength for two datasets and Bing/Huliu for the remaining two which is for both MMM events. Overall, the precision and recall for each dataset vary, where negative precision and recall has the closet range between each other from 0.70s to 0.80s with precision higher than recall. Neutral has range of up to 0.20 between precision and recall, with recall higher than precision and lastly positive has the widest difference with precision highest and recall lowest with an approximate difference of up to 0.50.

Top f-measure for each dataset sentiment category			
Dataset	**Negative**	**Neutral**	**Positive**
2015 MMM	Jockers family 0.65	Bing Huliu 0.69	Bing 0.31
2016 MMM	Jockers Rinker Sentimentr Jockers 0.56	Loughran MacDonald 0.71	Huliu 0.44
2016 Dover	Jockers Rinker 0.76	Loughran MacDonald Inquirer 0.48	SentiStrength 0.20
2016 Anti-Austerity	Syuzhet Jockers 0.35	Loughran MacDonald 0.77	SentiStrength 0.37

Table 27 Top F-measure for each sentiment category for MR2

The breakdown of dictionaries' strength based on sentiment categories shows how well they have performed across each sentiment categories. For example, for 2015 MMM "Huliu" performs best for neutral than positive and negative, and for most "Loughran MacDonald" performs best with neutral. Furthermore, for both Dover and AA "SentiStrength" performs best in the positive category as occurs twice.

MR2 agrees with MR1 that "Jockers Rinker" performs best with negative. In MR2 "SentiStrength" performs nearly the best for positive and "Loughran MacDonald" for neutral, but for MR1 for both neutral and positive "SentiStrength" performs best in each of these categories. Additionally, there is further agreement between MR1 and MR2 that 2015 MMM best performing dictionary for positive could be "Bing". Lastly, both MR1 and MR2 agreement "Slangsd" has the lowest f-measure with both "Socal Google" and Berkeley just as worse off. In section 6.5.3 we will explore both macro and micro precision, recall and F1.

6.5.3 Macro/ Micro Precision and Recall

Both Micro and Macro precision (defined in section 5.10), recall and F1 measure are used to evaluate the results for whether there is any class imbalance that impacts the strength of the outcome for the dictionary approach.

6.5.3.1 MR1 Results

Both Table 28 and Table 29 shows precision, recall and F1 scores for dictionaries in the top 3 and lowest positions for each dataset.

In Table 28 it can be seen that one or more Jockers family and "SentiStrength" dictionaries macro/micro F1 perform the best for both MMM events. This seems to be because of both a high micro/macro precision and recall, but precision is slightly higher in both of these categories which shows it has been overly aggressive in identifying positive cases, thus provides a lack of indication between the sentiment categories, which is not a good result.

The worst performing dictionaries are mostly “Berkeley”, “Socal Google” and “Senticnet” for F1 in both macro/micro categories.

MMM 2015	1	2	3	Lowest
Micro Precision	Bing, Huliu and Inquirer 0.81	Loughran Mcdonald 0.77	Vadar and Afinn 0.76	Berkeley 0.43 and Senticnet 0.44
Micro Recall	Jockers Rinker 0.67	Syuzhet and Sentimentr Jockers 0.66	Afinn and Vadar 0.57	Both Berkeley and Senticnet 0.28
Micro F-measure	Jockers Rinker, Syuzhet Jockers and Sentimentr Jockers 0.68	Vadar and Afinn 0.65	Huliu 0.64	Berkeley 0.34, both Socal Google and Senticent 0.39
Macro Precision	Vadar, Bing and Inquirer 0.78	Huliu 0.77	Afinn 0.74	Sentistrength 0.40 and Slangsd 0.42
Macro Recall	Jockers Rinker 0.61	Both Syuzhet and Sentimentr Jockers 0.60	Vadar 0.51	Socal Google 0.24, both Sentistrength and Berkeley 0.29
Macro F-measure	Same as above, but a score of 0.66	Vadar 0.62	Afinn 0.60	SentiStrength on 0.33, with Socal Google 0.34
MMM 2016				
Micro Precision	Bing and Inquirer 0.75	Huliu 0.74	Loughran Mcdonald and Sentistrength 0.73	Senticnet and Berkeley 0.36
Micro Recall	Sentistrength, Jockers Rinker and Syuzhet Jockers 0.58	Sentiment jockers 0.57	Afinn 0.56	Berkeley 0.24 and Senticnet 0.31
Micro F-measure	SentiStrength 0.65	Syuzhet Afinn 0.63	Vadar 0.62	Berkeley 0.29 and Socal Google fourth from bottom on 0.43
Macro Precision	Vadar 0.76	Bing 0.75	Huliu and Inquirer 0.74	Slangsd 0.37 and other common bottom ones higher up
Macro Recall	Jockers Rinker 0.58	Syuzhet Jockers 0.57	Sentimentr Jockers 0.56	Berkeley 0.28 and Socal Google 0.29
Macro F-measure	Jockers Rinker 0.63	Sentimentr jockers 0.62, Vadar 0.62 and Syuzhet jockers 0.62	Sentistrength and Afinn 0.61	Lowest are both Slangsd and Berkeley on 0.36

Table 28 Micro and Macro Averages for both MMM

In Table 29 it can be seen that “SentiStrength” is more dominate rather than Jockers family (in relation to both MMM events) for best dictionaries in macro/micro F1 performance except for macro in Dover which highest performers are both “Syuzhet Jockers” and

"Sentimentr Jockers". This seems to be because of a high micro/macro precision and lower recall scores for "SentiStrength" and "Jockers". The worst performing dictionaries are interchangeable between "Berkeley", "Socal Google", "Slangsd" and "Senticnet" for F1 in both macro/micro categories.

Dover 2016				
Micro Precision	SentiStrength and Loughran McDonald 0.88	Bing 0.87	Huliu 0.86	Senticnet 0.67, other common one higher up
Micro Recall	SentiStrength 0.53	Jockers Rinker and Syuzhet Jockers 0.52	Sentimentr Jockers 0.51	Socal Google 0.19, all Sentiword, Senticnet and Inuqirer 0.29
Micro F-measure	SentiStrength 0.66	Syuzhet Jockers 0.64, Jockers Rinker 0.64	Sentimentr jockers is 0.63	Lowest by far is Socal Google on 0.31
Macro Precision	Bing 0.85	SentiStrength 0.79	Huliu and Inquirer 0.78	Slangsd 0.49, other common ones higher up
Macro Recall	Jockers family 0.40	SentiStrength and Vadar 0.37	Afinn 0.34	Socal Google 0.16 and Slangsd 0.22
Macro F-measure	Syuzhet Jockers 0.52, Sentimentr Jockers 0.52	Jockers Rinker 0.51	SentiStrength and Vadar 0.50	Socal Google on 0.27
Anti-Austerity 2016				
Micro Precision	Loughran Mcdonald 0.78	SentiStrength 0.77	Bing and Huliu 0.76	Senticnet 0.36 and Berkeley 0.37
Micro Recall	SentiStrength 0.58	Stanford 0.57	Jockers family 0.48	Both Senticnet and Berkeley 0.27
Micro F-measure	SentiStrength 0.66,	Stanford 0.62	Sentimentr Jockers and Huliu is 0.58	Both Senticent and Berkeley on 0.31, Socal Google is fourth from bottom on 0.40
Macro Precision	Huliu 0.79	Bing and SentiStrength 0.78	Vadar and Afinn 0.76	Slangsd 0.37 lower down compared with the rest
Macro Recall	SentiStrength 0.54	Stanford 0.52	Jockers Rinker and Syuzhet Jockers 0.51	Socal Google 0.27 and Berkeley 0.30
Macro F-measure	SentiStrength 0.64	Jockers family are all on 0.60	Stanford, Huliu and Vadar 0.57	Slangsd 0.36 Socal google is third from bottom on 0.38.

Table 29 Micro and Macro Averages for Dover and Anti-Austerity

The Micro is higher in most instances, which indicates that those dictionaries perform well across every dataset results. Furthermore, the ones with a higher F1 Macro indicate that the classifier performs well for each individual class. However, for the 9 dictionaries (such as Jockers family, Vadar and Afinn across most datasets) in both Table 28 and Table 29 where Macro average is lower than the Micro average, there is poor metric performance on the smaller classes (Athar, 2014; Gate, 2019; Zhang, Wang & Zhao, 2015). Additionally, for the 3

dictionaries (such as Vadar and Jockers family for 2016 MMM) where the Micro average is lower than the macro average, there is poor performance on the larger classes.

The 2015 MMM Jockers family shows a higher Micro F-measure than Macro F-measure, which is not far behind. Jockers family tend to be in the top 3 of the other data-sets results, where in Dover it dominates 2nd and 3rd place. The number one position for Micro F-measure in the other three datasets is "SentiStrength" with both Macro and Micro top position for AA. Furthermore, other dictionaries appear once with no common pattern in each of the datasets' Macro/Micro F-measure. Additionally, Vadar has the same Micro and Macro F-measure in MMM 2016 with an equal f-measure of 0.62 which indicates an exact distribution of the scores or that the classifier has the same performance for all classes involved, thus the dictionary is well-balanced.

The Jockers family is top 3 for 2015 MMM and 2016 Dover, with the rank order being slightly different in the top 3, but the Micro F-measure shows a higher level of fluctuation than Macro F-measure specifically in Dover results. Furthermore, Jockers family is in top positions for both 2016 MMM and 2016 AA for Macro F-measure except that "SentiStrength" is top for AA and "Vadar" is in the top 3 for 2016 MMM. Macro F1 is higher than micro only for Jockers family and Vadar for both 2016 AA and 2016 MMM.

Overall, "SentiStrength" has scored to a high level across most datasets with a higher Micro F1 and lower Macro F1 except for 2015 MMM where SentiStrength is the lowest Macro F1 on 0.33 with a slightly higher Micro F1, but not included in the top 3 positions. As indicated in the results from both AA and 2016 MMM, in both Table 28 and Table 29 demonstrates only a few dictionaries have performed the best on each individual class, except Dover is different where Micro outperforms Macro by 0.12. This emphasises a good performance overall, but has some class imbalance. Additionally, shows 2015 MMM has both F1 scores as equally low, thus a poor performance of class distribution and on larger classes.

6.5.3.2 MR2 Results

Both Table 30 and Table 31 shows precision, recall and F1 scores for dictionaries in the top 3 and lowest positions for each dataset.

In Table 30 it can be seen that there are variety (includes Huliu, Vadar and Loughran MacDonald) of dictionaries that perform the best in macro/micro F1 for both MMM events. Therefore, there is no definitive best performer in both macro and micro F1. These dictionaries F1 macro/micro results are due to a high micro/macro precision and recall, but precision is slightly higher in both of these categories which shows it has been overly aggressive in identifying positive cases, thus provides a lack of indication between the sentiment categories, which is not a good result. The worst performing dictionaries are mostly "Berkeley" and "Senticnet" for F1 only for micro precision and recall, but for macro

precision interchangeable between “SentiStrength”, “Stanford” and “Slangsd”. However, macro recall “Socal Google” is consistently the worst performer on the lowest score.

MMM 2015	1	2	3	Lowest
Micro Precision	Bing and Inquirer 0.62	Huliu 0.61	Loughran MacDonald 0.59	Senticnet and Berkeley 0.26
Micro Recall	Syuzhet Jockers and Jockers Rinker 0.66	Sentimentr Jockers 0.65	Afinn 0.64	Berkeley 0.26 Senticnet 0.32
Micro F-measure	Huliu 0.61	Bing and Afinn 0.60	Loughran MacDonald 0.59	Berkeley 0.26 Senticnet 0.29
Macro Precision	Vadar and Bing 0.58	Inquirer and Huliu 0.57	Syuzhet NRC 0.55	SentiStrength 0.33 Stanford 0.34
Macro Recall	Syuzhet Jockers and Jockers Rinker 0.63	Senitmentr Jockers 0.62	Vadar 0.56	Socal Google 0.30 Syuzhet NRC 0.34
Macro F-measure	Vadar and Syuzhet Jockers 0.57	Sentimentr Jockers and Jockers Rinker 0.56	Afinn 0.55	SentiStrength 0.35 Stanford 0.36
MMM 2016				
Micro Precision	Inquirer 0.61	Bing, Huliu and Loughran MacDonald 0.59	Socal Google 0.57	Senticnet 0.24 Berkeley 0.25
Micro Recall	SentiStrength 0.62	Afinn, Jockers Rinker and Loughran MacDonald 0.61	Syuzhet Jockers, Vadar and Sentimentr Jockers 0.60	Berkeley 0.25 Senticnet 0.27
Micro F-measure	Loughran MacDonald 0.60	Huliu, Bing, SentiStrength, Inquirer and Afinn 0.58	Vadar 0.54	Berkeley 0.25 Senticnet 0.27
Macro Precision	Vadar 0.58	Huliu and Bing 0.57	Inquirer 0.56	Slangsd 0.30 Stanford 0.34
Macro Recall	Jockers Rinker 0.63	Syuzhet Jockers 0.62	Sentimentr Jockers 0.61	Socal Google 0.34 Berkeley 0.35
Macro F-measure	Vadar 0.58	Jockers Rinker and Sentimentr Jockers 0.57	Syuzhet Jockers 0.56	Slangsd 0.35 Berkeley 0.37

Table 30 Micro and Macro Averages for both MMM for MR2

In Table 31 a similar pattern emerges with Table 30 where again there are variety (includes SentiStrength, Jockers Rinker and Loughran MacDonald, but a greater number of dictionaries on joint score for AA) of dictionaries that perform the best in macro/micro F1 for both events. Therefore, again there is no definitive best performer in both macro and micro F1. Similarly, to both MMM events, these dictionaries F1 macro/micro results are due

to a high micro/macro precision and recall with precision is slightly higher, thus provides a lack of indication between the sentiment categories, which is not a good result.

The worst performing dictionaries for F1 are mainly “Socal Google” and “Senticnet” for both Dover and AA. The micro/macro precision and recall are inconsistent on its lowest performing dictionaries (interchangeable between “SentiStrength”, “Stanford”, “Senticnet” and “Slangsd” compared to previous results for MR1. These poor results are reflected in the MR2 Dover and AA results show micro is slightly higher than macro for the lowest dictionaries.

Dover 2016	**1**	**2**	**3**	**Lowest**
Micro Precision	Loughran MacDonald 0.70	Inquirer 0.67	Sentiment NRC, SentiStrength and Syuzhet NRC 0.66	Senticnet 0.44 SentiWord 0.46
Micro Recall	SentiStrength 0.56	Jockers Rinker 0.53	Syuzhet Jockers and Afinn 0.52	Socal Google 0.23 Senticnet 0.28
Micro F-measure	SentiStrength 0.60	Afinn 0.57	Jockers Rinker, Vadar and Sentimentr Jockers 0.56	Socal Google 0.33 Senticnet 0.34
Macro Precision	Inquirer 0.64	Sentiment NRC 0.62	Syuzhet NRC, Bing, Huliu and Vadar 0.60	Slangsd 0.41 Stanford 0.45
Macro Recall	Jockers Rinker 0.46	Sentimentr Jockers and Syuzhet Jockers 0.45	Vadar and SentiStrength 0.43	Socal Google 0.21 Slangsd 0.26
Macro F-measure	Jockers Rinker and Sentimentr Jockers 0.51	Syuzhet Jockers and Vadar 0.50	SentiStrength and Afinn 0.49	Socal Google 0.30 Slangsd 0.32
Anti-Austerity 2016				
Micro Precision	Loughran MacDonald 0.58	Socal Google and Bing 0.51	Inquirer 0.50	Senticnet 0.14 Berkeley 0.15
Micro Recall	Stanford 0.70	SentiStrength and Loughran MacDonald 0.68	Bing 0.62	Senticnet 0.21 Berkeley 0.22
Micro F-measure	Loughran MacDonald 0.63	Bing 0.56	SentiStrength 0.55	Senticnet 0.17 Berkeley 0.18
Macro Precision	Bing 0.48	Huliu and Afinn 0.47	Vadar and Inquirer 0.46	Slangsd 0.22 SentiWord 0.32
Macro Recall	Stanford 0.59	SentiStrength 0.58	Syuzhet Jockers and Jockers Rinker 0.57	Socal Google 0.36 Slangsd 0.42
Macro F-measure	Huliu, Bing and Syuzhet Jockers 0.50	Sentimentr Jockers, Afinn, Jockers Rinker and SentiStrength 0.49	Syuzhet NRC, Combined Dictionary and	Slangsd 0.29 Socal Google 0.38 SentiWord 0.38

			Sentiment NRC 0.47	

Table 31 Micro and Macro Averages for Dover and Anti-Austerity for MR2

The majority of MR2 Micro F1 scores are higher than Macro F1 scores with most being in a similar range to each other.

Although the Micro f-measure again has more scores higher than Macro f-measure, but overall MR1 has higher Micro/Macro F1 scores compared to MR2, which shows MR1 has less imbalance. MR2's Micro dominates with all datasets results, but generally lower than MR1, therefore, there is a higher level of imbalance on larger classes. There are 6 instances (Vadar and mainly Jockers Family) where Macro is higher than Micro with a small difference in range for both MMM results. This is similar to MR1, which indicates a poor metric performance on smaller classes.

Both MMM events for Micro F1s have a common agreement that both Bing and Afinn are consistently strong performers in second place, with 2016 MMM showing a slightly lower score. Loughran MacDonald is third strongest performer for 2015 MMM, but first for 2016 MMM with a marginal difference between second/ third top positions. Additionally, both MMM for Micro F1 agree that both Berkeley and Senticnet have the worst score, but for lowest Macro there are a list of inconsistent dictionaries listed, thus no single definitive dictionary can be chosen. The highest Macro F1 is strongest with Vadar for both MMM, but 2015 MMM Vadar is joint with Syuzhet Jockers. For second strongest both MMM have the exact same dictionaries of Sentimentr Jockers and Jockers Rinker, but for third there is no agreement on the other best performing dictionary.

Both Anti-Austerity and Dover has the worst scores for macro/micro F1 compared to both MMM events, but Anti-Austerity Loughran MacDonald performs the best with the highest micro f-measure except for Dover where it is highly changeable for the top 3 dictionaries. The datasets results show there is common agreement that Senticnet has the worst F1 score except for Dover which outlines it as 2nd lowest. Dover has less in agreement with the other datasets which is due to the higher level of negativity. MR1 has a higher level of agreement between the top 3/ lowest places than MR2 shows MR1 has less imbalance due to the higher F1 scores and narrower range between both micro/macro F1.

Both micro and macro precision, recall and F1 have been explored for each dictionary to determine the strength of the results. Section 6.5.5 will apply the dictionaries results in the machine learning phase to predict the sentiment for new series of tweets based on a sample of the UK demonstration tweets that have been classed as identified in the automated coded results.

6.5.4 Dictionary: Machine Learning Results

The input data for train and validation will be split 80/20 from 1500 rows of 19 dictionaries sentiment classification results and manual classification for each dataset, equalling 20 columns. However, for both MR1 grouped and MR2 grouped the trainset will combine each dataset, randomise it and split it into 4800 trainset and 1200 for validation. These dictionaries results are compared to the manual classification separately for MR1 and MR2 based on the tweet's sentiment labelled either 'negative', 'neutral' or 'positive' to build models to predict sentiment for MR1 and MR2 tweets. Additionally, the validation data will be fed against several different models (e.g. Random Forest, Maximum Entropy, Support Vector Machine (SVM), Bagging, Decision Tree and Naïve Bayes) from the automated coded (relevant) data from each of the four separate datasets and also a combined version of all four datasets. These results, will again consist of precision, recall and the f-measure to determine the strength of the outcome.

To run these algorithms R is used to access the first five algorithms within "RTextTools" package, whilst Naïve Bayes can be only applied from the "Caret" package. The default settings for each algorithm are applied in this machine learning process. However, in general 'RTextTools' is limited in its customisation of the algorithm settings when compared to 'Caret'.

6.5.4.1 MR1 Results

The tables in this section provides a summary of the MR1 results for the various algorithms. In all the tables accuracy of the predictions is largely correct where for each case it shows that negative and then neutral have the highest overall accuracy. For instance, in Table 32, 2016 Dover, there are 238 correct and 62 incorrect with the highest accuracy score of 0.79. The classifier has correctly predicted 238 negative and incorrectly predicted 57 neutral and 5 positives. The accuracy is 0.79, but it appears that the classifier is not as capable to recognising the other sentiment categories. This could be due to the sample of train data leaning towards negativity than a balance across each class. Despite this, this dataset tweets are mainly negative. The accuracy indicator has proven less helpful; therefore, it will no longer be discussed in the results.

In Table 32 all datasets F1 measure varies from each sentiment class where negative has the highest scores from 0.73 to 0.89, neutral from 0.60 to 0.66 (apart from Dover as NA due no neutrals identified by Naïve Bayes), and positive from 0.4 to 0.58 except for both Dover and 2015 MMM is NA due to no positives identified. Naïve Bayes performed well across each sentiment category for both 2016 MMM and AA that tend to have more occurrences of positives. For 2015 MMM and Dover there was much higher negative and neutral in the sample rather than positives, which seems to be why zero positives have been classified.

In Table 32 the Anti-Austerity trainset contained a higher number of positives tweets, despite this it has fewer positive tweets than 2016 MMM. It cannot be compared to 2015 MMM and Dover as these contained higher levels of negative and neutral tweets in its sample. Despite the fact that AA is a more balanced sample between the sentiment categories, it contains the most incorrectly classified tweets. This specific dataset's balance may have caused the classifier to find it more difficult to determine the correct classification for each category. In the future, it may be that more train data is required to help the algorithms to detect the signal better.

	CARET (R package)						
		Predicted Classes			**Predicted Classes**		
	Prediction: Naïve Bayes (MMM 2015)				**Naïve Bayes (MMM 2016)**		
		Negative	Neutral	Positive	Negative	Neutral	Positive
Actual Classes	**Negative**	165	45	14	112	45	8
	Neutral	16	60	0	26	78	2
	Positive	0	0	0	2	11	16
	Precision	0.74	0.79	NA	0.68	0.74	0.55
	Recall	0.91	0.57	NA	0.8	0.58	0.62
	F1	0.82	0.66	NA	0.73	0.65	0.58
	Accuracy	0.75			0.69		
	Prediction: Naïve Bayes (Dover)				**Naïve Bayes (Anti-Austerity)**		
Actual Classes	Negative	238	57	5	127	27	0
	Neutral	0	0	0	41	65	1
	Positive	0	0	0	10	19	10
	Precision	0.79	NA	NA	0.83	0.61	0.26
	Recall	1	0	0	0.72	0.59	0.91
	F1	0.89	NA	NA	0.77	0.60	0.4
	Accuracy	0.7933			0.6733		

Table 32 Naive Bayes results for all datasets

In Table 33 2015 MMM F1 varies from 0.5 to 0.67 with Tree the worst and Bagging the strongest. Bagging F1 is higher than Naïve Bayes as more has been correctly classified across each sentiment category. Forest, Tree and Bagging identify positives albeit little, but is an improvement on the result for Naïve Bayes.

	MMM 2015					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	151	29	1	148	30	3
Neutral	37	68	0	32	73	0
Positive	6	4	4	4	5	5
Precision	0.75			0.70		

Recall	0.59			0.63		
F1 Score	0.63			0.65		
Accuracy	0.74			0.75		
	Support Vector Machine			**Bagging**		
Negative	151	30	0	149	31	1
Neutral	37	68	0	39	66	0
Positive	6	8	0	4	4	6
Precision	0.47			0.76		
Recall	0.49			0.63		
F1 Score	0.48			0.67		
Accuracy	0.73			0.74		
	Tree					
Negative	158	23	0	**Precision**	0.50	
Neutral	37	68	0	Recall	0.51	
Positive	12	2	0	F1 Score	0.50	
				Accuracy	0.75	

Table 33 2015 MMM results of other machine learning algorithms

In Table 34, 2016 MMM the highest F1 score is Max Entropy of 0.68 (which is slightly higher than Naïve Bayes and is third place) and the lowest of 0.59 for SVM although it seems due to both high precision and high recall with precision being slightly higher. Overall, 2016 MMM has a slightly lower precision, recall and F1 compared to 2015 MMM which has a wider gap where precision is higher than recall.

	MMM 2016					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	111	28	1	99	39	2
Neutral	34	93	7	21	105	8
Positive	13	6	7	8	4	14
Precision	0.63			0.69		
Recall	0.58			0.68		
F1 Score	0.60			0.68		
Accuracy	0.70			0.73		
	Support Vector Machine			**Bagging**		
Negative	109	30	1	101	38	1
Neutral	34	95	5	33	95	6
Positive	12	8	6	11	8	7
Precision	0.64			0.64		
Recall	0.57			0.58		
F1 Score	0.59			0.60		
Accuracy	0.70			0.70		
	Tree					
Negative	108	30	2	**Precision**	0.69	
Neutral	32	96	6	Recall	0.66	
Positive	5	8	13	F1 Score	0.67	
				Accuracy	0.72	

Table 34 2016 MMM results of other machine learning algorithms

In Table 35, the Dover performance for F-measure's best result is Tree at 0.48 (is higher than Naïve Bayes which is the worst performer across the sentiment categories) and Max Entropy at 0.44 with the worst SVM at 0.37 although it seems due to both high precision and high recall with precision being slightly higher. Overall, Dover has produced the worst F1 results which shows each algorithm struggled which is due to Dover having the highest level of negativity in its sample data.

	DOVER 2016					
	Accuracy					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	231	7	0	235	3	0
Neutral	43	14	0	41	16	0
Positive	5	0	0	5	0	0
Precision	0.50			0.56		
Recall	0.41			0.42		
F1 Score	0.42			0.44		
Accuracy	0.82			0.84		
	Support Vector Machine			**Bagging**		
Negative	228	10	0	225	13	0
Neutral	48	9	0	42	15	0
Positive	5	0	0	5	0	0
Precision	0.43			0.46		
Recall	0.37			0.40		
F1 Score	0.37			0.41		
Accuracy	0.79			0.80		
	Tree					
Negative	228	10	0	**Precision**	0.52	
Neutral	32	25	0	**Recall**	0.47	
Positive	5	0	0	**F1 Score**	0.48	
				Accuracy	0.84	

Table 35 2016 Dover results of other machine learning algorithms

In Table 36, AA results for its strongest F1 performance is Tree on 0.65 (which is higher than Naïve Bayes which has less correctly classified that would sit be third from bottom) and Max Entropy on 0.63 with the worst being Bagging on 0.52 although it seems due to both high precision and high recall with recall being slightly higher, whereas the other datasets had marginal gain for precision. Overall, AA has produced the third highest F1 results behind 2015 MMM in first and 2016 MMM in second.

	Anti-Austerity 2016					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	135	38	5	130	45	3
Neutral	33	70	8	27	76	8
Positive	4	5	2	1	3	7

Precision	0.51			0.61		
Recall	0.52			0.68		
F1 Score	0.51			0.63		
Accuracy	0.69			0.71		
	Support Vector Machine			**Bagging**		
Negative	136	37	5	132	39	7
Neutral	30	73	8	28	73	10
Positive	4	5	2	7	2	2
Precision	0.52			0.51		
Recall	0.53			0.53		
F1 Score	0.52			0.52		
Accuracy	0.70			0.69		
	Tree					
Negative	133	42	3	**Precision**	0.62	
Neutral	24	79	8	**Recall**	0.7	
Positive	0	4	7	**F1 Score**	0.65	
				Accuracy	0.73	

Table 36 2016 Anti-Austerity results of other machine learning algorithms

Tree and Max Entropy has the highest F1 for most datasets except for 2015 MMM which Bagging is the best with Max Entropy in second place and the worst is Tree. Max Entropy could be considered the consistently higher performer across each of the datasets. Max Entropy has correctly classified more positivity on most occasions, which has contributed to the good performance. However, for Dover no algorithm identified any positivity and Max Entropy on this occasion had fewer correct classified for each sentiment category, which is why is second behind Tree. Furthermore, Bagging had slightly more correctly classified, but more specifically in the positive category by +1 which is why it just sits above Max Entropy. In strongest performing results precision and recall are nearly similar in score, but on most occasions, precision is slightly higher except for AA which higher recall over precision.

In the fitting of several models, the results (refer to Table 139 to Table 142 in appendix 10.12.1.1) are reflective of the results produced so far above, and re-affirmed in Table 143 in appendix 10.12.1.1 when all trainsets from each dataset are combined to 4800 and randomised then placed against the validation set of 1200. The numbers show a similar split by proportion when compared to the first four datasets standalone, showing negative and neutral are higher than positive. The models are tested against the set of automated relevant tweets to predict their sentiment. In Table 37, in all the datasets in the predicted sentiment categories, the classifiers that are highest are negative and neutral, with negative with the majority on most occasions except for Anti-Austerity, in which neutral is dominant. Anti-Austerity has the highest positive by far compared to the other datasets, with Dover on zero. These results share similarity with the model results, with negative and neutral being the highest. Overall, the sample of data requires widening to include more positive tweets as this may change the overall result in the future.

2015 MMM Machine Learning Results						
SENTIMENT_LABEL	**MAXENTROPY_LABEL**	**SVM_LABEL**	**FORESTS_LABEL**	**BAGGING_LABEL**	**TREE_LABEL**	**Naïve Bayes (CARET)**
Negative	14843	16823	17053	16709	19163	21618
Neutral	12528	12458	10928	11178	10257	7533
Positive	2049	139	1439	1533	NA	269
2016 MMM Machine Learning Results						
Negative	6489	7472	7627	7379	7491	8091
Neutral	7726	7060	6823	7043	6931	5711
Positive	1276	959	1041	1069	1069	1689
2016 DOVER Machine Learning Results						
Negative	2814	2790	2792	2758	2576	3174
Neutral	360	384	382	416	598	NA
Positive	NA	NA	NA	NA	NA	NA
Anti-Austerity 2016 Machine Learning Results						
Negative	9792	11049	11190	11225	9365	9682
Neutral	15813	15001	14860	13974	16115	12245
Positive	4358	3913	3913	4764	4483	8036

Table 37 Dictionary Approach - MR1 Machine Learning Results tested on automated relevant tweets from each dataset

6.5.4.1.1 MR1 Grouped (Combined)

As iterated in section 6.5.4 the train set is 4800 and validation is 1200 first, which is same 80/20 split, then is tested against automated coded (relevant) datasets to predict their sentiment.

In Table 38 shows Max Entropy has the strongest F1 score of 0.66, but by a small margin of +0.01 in front of Forest. The worst performing algorithm is Naïve Bayes which has less correctly classified behind both SVM and Bagging. Although it seems due to both high precision and high recall with precision being slightly higher. All the results show that negativity has the highest correctly classified, and neutral is further behind in proportion than what is displayed in the results above on an individual level.

	Grouped Trainset Algorithm Results					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	569	99	18	569	99	18
Neutral	141	286	17	135	292	17
Positive	17	19	34	23	13	34
Precision	0.66			0.66		
Recall	0.65			0.66		
F1 Score	0.65			0.66		
Accuracy	0.74			0.75		
	Support Vector Machine			**Bagging**		
Negative	579	94	13	573	100	13
Neutral	161	272	11	156	277	11
Positive	25	18	27	25	18	27
Precision	0.67			0.66		
Recall	0.61			0.62		
F1 Score	0.64			0.64		
Accuracy	0.73			0.73		
	Tree			**Naïve Bayes**		
Negative	531	133	22	624	236	36
Neutral	126	299	19	46	198	1
Positive	5	28	37	16	10	33
Precision	0.64			0.70	0.81	0.56
Recall	0.66			0.91	0.45	0.47
F1 Score	0.65			0.79	0.58	0.51
Accuracy	0.72			0.71		

Table 38 MR1 Grouped (Combined) Algorithm Results

In Table 39, the dictionary machine learning results based on new test data against the model shows that it is reflective of the above results as are similar by proportion in Table 33 to Table 37, for example, Anti-Austerity shows a higher neutral count than negative, but most are negative and neutral than positive. The grouped set highest correctly classified performs consistently well for each algorithm except for Tree, with neutral being the largest. Additionally, the correctly classified distance between negative and neutral is mainly between 6,000 to 10,000 but Naïve Bayes is the exception on 45,000.

SENTIMENT_LABEL	MAXENTROPY_LABEL	SVM_LABEL	FORESTS_LABEL	BAGGING_LABEL	TREE_LABEL	Naïve Bayes (CARET)
2015 MMM Datasets (grouped-subset) - Machine Learning Results						
Negative	16764	17658	17028	17231	14727	21347
Neutral	11142	10534	10878	10961	12903	6706
Positive	1514	1228	1514	1228	1790	1367
2016 MMM Datasets (grouped-subset) - Machine Learning Results						
Negative	8158	8536	8168	8411	7491	10385
Neutral	6419	6269	6409	6394	6931	4169
Positive	914	686	914	686	1069	937
2016 Dover Datasets (grouped-subset) - Machine Learning Results						
Negative	2159	2239	2192	2211	2024	2562
Neutral	890	835	857	863	999	494
Positive	125	100	125	100	151	118
2016 AA Datasets (grouped-subset) - Machine Learning Results						
Negative	11979	12761	11469	12456	9365	19015
Neutral	13940	13840	14450	14145	16115	7530
Positive	4044	3362	4044	3362	4483	3418
2016 Grouped-Subset with Grouped Unseen - Machine Learning Results						
Negative	39060	41194	38857	40309	33607	53309
Neutral	32391	31478	32594	32363	36948	18899
Positive	6597	5376	6597	5376	7493	5840

Table 39 MR1 Grouped (Combined) Algorithm Result on tested on automated relevant tweets from each dataset

6.5.4.2 MR2 Results

The tables in this section provides a summary of the MR2 results for the various algorithms. In Table 40 shows that most of the time the predictions for correctly classifying the data is largely correct and for each demonstration show both negative and then neutral have the highest count except for Anti-Austerity which has no negatives with the majority being neutral.

In Table 40, for 2015 MMM, the Naïve Bayes (NB) classifier correctly predicted 203 (+27 more than MR1), and the breakdown is 'negative' for 74 (+91) cases, but the remaining 16 (+29) for neutral and 11 (+3) for positive are incorrectly classified. Additionally, it correctly predicted 129 (-69 less than MR1) neutral, but incorrectly predicted 18 (-2) negative cases and for positive there is 7 (-7). There is only 0s for the positive category. In 2016 MMM, there are 185 (+21) classified correct, with 30 (+82) negative, both neutral 20 (+58) and 5 (+11) positives (+69) are incorrect. NB classifier for 2015 MMM has more correct than 2016 MMM, but 2016 has capability to recognise a positive category, which has a F1 score 0.40, than NA for 2015, but overall 2015 MMM has stronger F1 scores for negative and neutral.

CARET (R package)						
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Naïve Bayes (MMM 2015)			**Naïve Bayes (MMM 2016)**		
Negative	74	61	11	30	20	5
Neutral	18	129	7	50	140	23
Positive	0	0	0	3	14	15
Precision	0.51	0.84	NA	0.55	0.68	0.47
Recall	0.80	0.68	0	0.36	0.81	0.35
F1 Score	0.62	0.75	NA	0.44	0.72	0.40
Accuracy	0.68			0.62		
	Naïve Bayes (Dover)			**Naïve Bayes (Anti-Austerity)**		
Negative	207	89	4	0	0	0
Neutral	0	0	0	40	229	15
Positive	0	0	0	0	10	6
Precision	0.69	NA	NA	NA	0.81	0.38
Recall	1	0	0	0	0.96	0.29
F1 Score	0.82	NA	NA	NA	0.88	0.32
Accuracy	0.69			0.78		

Table 40 Naive Bayes results

In Table 40 2016 Dover there are overall 207 (-31) correctly classified and 93 (+31) incorrectly. The classifier has correctly predicted 207 (-31) negative and incorrectly predicted 89 (-32) neutral and 4 (+1) positive. The algorithm's identifying positive and neutral is lesser due to the sample train data being more negative. This is supported by MR1 Dover which had shared similar results. In Table 40, 2016 Anti-Austerity NB classifies 235 (-33) correctly and 65 (+33) are incorrect. The Anti-Austerity trainset contained the highest

number of neutral tweets and is second place for total count of positives with 2016 MMM which is similar to MR1 results. The Anti-Austerity trainset is not evenly balanced as shown in the MR1 result. The most evenly balanced dataset in Table 40 is 2016 MMM, but it contains the most incorrectly classified tweets between the categories, which is exactly the same as the MR1 Anti-Austerity result. This further supports that NB finds it more difficult to determine the correct classification for each category. This problem may be improved upon with more train data to balance the sentiment categories, which is where further experimentation would need to be conducted in the future.

In Table 41, the algorithms' F1 scores are significantly lower than the MR1 results. The strongest algorithm is Max Entropy of 0.53 with both Bagging and Forest on 0.50 as it has more correctly classified and contained positives in its sentiment category. Whereas, both SVM and Tree are in the 40s due to less incorrect results and containing no positives. These results shows that the MR2 2015 MMM algorithms agree with MR1 on how the highest F1 score differs from highest to lowest.

	MMM 2015					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	67	25	0	63	29	0
Neutral	63	127	0	43	147	0
Positive	6	10	2	4	12	2
Precision	0.76			0.78		
Recall	0.50			0.52		
F1 Score	0.50			0.53		
Accuracy	0.65			0.71		
	Support Vector Machine			**Bagging**		
Negative	59	33	0	69	23	0
Neutral	57	133	0	65	125	0
Positive	7	11	0	6	10	2
Precision	0.41			0.76		
Recall	0.45			0.51		
F1 Score	0.42			0.50		
Accuracy	0.64			0.65		
	Tree					
Negative	83	9	0	**Precision**	0.44	
Neutral	82	108	0	**Recall**	0.49	
Positive	8	10	0	**F1 Score**	0.44	
				Accuracy	0.64	

Table 41 2015 MMM results of other machine learning algorithms

In Table 42, 2016 MMM is in the mid-60s, similar to 2015 MMM, but is consistently lower for each algorithm shared with MR1 results. However, MR1 for 2016 MMM algorithm has a F1 score more evenly balanced between the algorithms than MR2, as it has a higher level of agreement with the dictionary classification. The highest F1 score of 0.53 (+15) for Max Entropy and the lowest F1 score is Tree on 0.39. MR1 Max Entropy has the highest F1 score,

which is similar to the MR2 results. This further supports MR1 has a stronger set of results compared to MR2 F1 scores.

	MMM 2016					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	33	47	3	29	52	2
Neutral	26	144	4	15	152	7
Positive	5	27	11	5	25	13
Precision	0.60			0.61		
Recall	0.50			0.51		
F1 Score	0.52			0.53		
Accuracy	0.63			0.65		
	Support Vector Machine			**Bagging**		
Negative	3	80	0	34	46	3
Neutral	1	172	1	22	148	4
Positive	1	36	6	12	26	5
Precision	0.69			0.53		
Recall	0.39			0.46		
F1 Score	0.36			0.46		
Accuracy	0.60			0.62		
	Tree					
Negative	80	3	0	**Precision**	0.40	
Neutral	167	7	0	**Recall**	0.43	
Positive	29	14	0	**F1 Score**	0.39	
				Accuracy	0.29	

Table 42 2016 MMM results of other machine learning algorithms

In Table 43, the 2016 Dover results for MR2 are considerably lower than MR1, as MR1 has three algorithms higher or equal to Naïve Bayes. MR2 is not in agreement with MR1's view as all the algorithms being much lower. The highest F1 score is Max Entropy of 0.47 and the lowest is SVM on 0.38 of which MR2 agrees with MR1 that SVM is the worst performing dictionary for Dover results.

	DOVER 2016					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	183	24	0	188	19	0
Neutral	58	31	0	47	42	0
Positive	2	2	0	2	2	0
Precision	0.43			0.49		
Recall	0.41			0.46		
F1 Score	0.41			0.47		
Accuracy	**0.71**			**0.77**		
	Support Vector Machine			**Bagging**		
Negative	188	19	0	185	20	2
Neutral	67	22	0	61	26	2
Positive	2	2	0	2	2	0
Precision	0.41			0.43		

Recall	0.39			0.39	
F1 Score	0.38			0.40	
Accuracy	**0.70**			0.70	
	Tree				
Negative	196	11	0	**Precision**	0.48
Neutral	62	27	0	**Recall**	0.42
Positive	3	1	0	**F1 Score**	0.42
				Accuracy	0.74

Table 43 2016 Dover results of other machine learning algorithms

In Table 44, the 2016 Anti-Austerity F1 scores for all algorithms in 0.30s, which is a poor set of results in comparison MR1 results. The highest F1 score of 0.35 Table 44, but Naïve Bayes has the strongest F1 score of 0.40. The lowest F1 score is SVM of 0.30, which is not the same as MR1 which was Bagging. This Dover result agrees SVM is the lowest with both MR1 and MR2.

	Anti-Austerity 2016					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	1	39	0	0	40	0
Neutral	11	227	1	0	239	0
Positive	1	19	1	0	20	1
Precision	0.46			0.60		
Recall	0.34			0.35		
F1 Score	0.33			0.33		
Accuracy	0.76			0.13		
	Support Vector Machine			**Bagging**		
Negative	0	40	0	2	38	0
Neutral	0	239	0	14	216	9
Positive	0	21	0	1	19	1
Precision	0.27			0.37		
Recall	0.33			0.35		
F1 Score	0.30			0.35		
Accuracy	0.13			**0.73**		
	Tree					
Negative	0	40	0	**Precision**	0.27	
Neutral	0	239	0	**Recall**	0.33	
Positive	0	21	0	**F1 Score**	0.30	
				Accuracy	0.13	

Table 44 2016 Anti-Austerity results of other machine learning algorithms

MR2 results show Max Entropy is consistently the best performing dictionary except for Anti-Austerity where Naïve Bayes was on top. The highest F1 score of 0.65 is for 2016 MMM, with the others behind by far with 2015 MMM of 0.53, Dover of 0.47 and lowest is Anti-Austerity of 0.40. Additionally, the lowest is SVM for 2015 MMM and Anti-Austerity, which is followed by Tree for 2016 MMM and Naïve Bayes for Dover. 2016 MMM had the highest F1 scores whereas the remaining results are much lower in 0.50s or 0.40s. MR2 mostly agrees with MR1 that Max Entropy is the best algorithm to correctly classify the data. MR1 has higher F1 scores for every dataset when compared with MR2, but there are

similarities with Max Entropy being highest and SVM being the lowest. Moreover, MR2 showed negative and neutral count highest, and any algorithm that assigned any positivity tends tend to have a higher F1 score compared to those without, which is supported by MR1 results. Furthermore, 3 out of 4 datasets showed neutral count as majority with the exception of Dover being negative.

MR2 does contain some unusual results, mainly because 2016 MMM and Anti-Austerity have scores that are very low and some that are very high in the same dataset, whereas MR1 scores remains consistent. These strange results led to running the same dictionary machine learning approach, but again it produced the same low and high scores. This would need further testing in the future to understand where the process could be improved to enhance the F1 scores and bring similar consistency to the algorithms results.

In the fitting of several models, the results (refer to Table 145 and Table 148 in appendix 10.12.2) are reflective of the results produced with neutral and negative highest count, with neutral with majority with exception with Dover being negative. This is re-affirmed in Table 149 in appendix 10.12.2 when all trainsets from each dataset are combined to 4800 and randomised then placed against the validation set of 1200. The numbers show a similar split by proportion when compared to the first four datasets standalone, showing negative and neutral are higher than positive, but there is an exception where Tree leads by two over neutral.

The train/validation models are tested against the set of automated coded (relevant) tweets to predict their sentiment. In Table 45, the classifiers predictions that are highest are negative and neutral, with negative with the majority for Dover, but for the rest of the demonstrations results the dominant category is neutral. Anti-Austerity has the highest positive total count by far compared to the other datasets, with Dover on zero, which is the same as MR1. These results share similarity with the model results, with negative and neutral being the highest. The algorithms with 1 or more correctly classified positive appeared to have a highest F1 score, so in Table 45 the algorithms with most positivity, some neutral and negativity are both MaxEnt and Bagging which is similar to MR1. These could be considered the strongest performers based on this theory for both MR1 and MR2. Overall, the sample of data requires widening to include more positive tweets as this may change the overall result in the future.

SENTIMENT_LABEL	MAXENTROPY_LABEL	SVM_LABEL	FORESTS_LABEL	BAGGING_LABEL	TREE_LABEL	Naïve Bayes (CARET)
2015 MMM Machine Learning Results						
Negative	8018	10506	10712	11430	13344	11949
Neutral	20809	18914	17671	16923	16076	17471
Positive	593	NA	1037	1067	NA	NA
						Accuracy
						0.6767
2016 MMM Machine Learning Results						
Negative	2723	221	3564	3694	NA	2607
Neutral	11690	14960	11055	10945	14146	10989
Positive	1078	310	872	852	1345	1895
						Accuracy
						0.6167
2016 Dover Machine Learning Results						
Negative	2272	2386	2377	2310	2556	3174
Neutral	902	788	797	838	618	NA
Positive	NA	NA	NA	26	NA	NA
						Accuracy
						0.69
2016 AA Machine Learning Results						
Negative	NA	NA	530	1330	NA	NA
Neutral	28942	29963	28558	27213	29963	25916
Positive	1021	NA	875	1420	NA	4047
						Accuracy
						0.7833

Table 45 Dictionary Approach - MR2 Machine Learning Results for New Data

6.5.4.2.1 MR2 Grouped (Combined)

As iterated in section 6.5.4 the train set is 4800 and validation is 1200 first, which is same 80/20 split, then second is tested against automated relevant datasets to predict their sentiment.

In Table 46 shows Naïve Bayes with the highest F1 score of 0.53, whereas the other algorithms are all in the 0.40s. Naïve Bayes has 762 correctly classified and is more evenly spread compared with the other results, which could be reason it has a higher F1 score. MR1 grouped results have a higher F1 score, and Max Entropy has the highest F1 score of 0.66.

	MR2 Grouped (Combined) Datasets					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	239	173	7	195	221	3
Neutral	190	494	12	137	551	8
Positive	12	65	8	9	71	5
Precision	0.50			0.51		
Recall	0.46			0.44		
F1 Score	0.46			0.44		
Accuracy	0.62			0.63		
	Support Vector Machine			**Bagging**		
Negative	231	181	7	238	174	7
Neutral	184	500	12	194	490	12
Positive	13	64	8	14	63	8
Precision	0.50			0.50		
Recall	0.45			0.45		
F1 Score	0.46			0.46		
Accuracy	0.62			0.61		
	Tree			**Naïve Bayes**		
Negative	321	98	0	304	238	25
Neutral	259	437	0	104	439	41
Positive	21	64	0	11	19	19
Precision	0.42			0.54	0.75	0.39
Recall	0.47			0.73	0.63	0.22
F1 Score	0.44			0.62	0.69	0.28
Accuracy	0.63			0.64		

Table 46 MR2 Grouped (Combined) Algorithm Results

In Table 47, the dictionary machine learning results based on new test data against the models shows that it is reflective in above results and the initial models output in Table 149 in appendix 10.12.2. Table 46 has a similar proportion to Table 47 algorithms results, as all the results show neutral and negative with highest count, but neutral has the largest consistency for every algorithm except Tree for Dover results that shows negative as the highest by 690. However, for MR1 the negative category has the majority proportion, which has more than likely determined the category for new data unless it has a more even spread across each sentiment class.

SENTIMENT_LABEL	MAXENTROPY_LABEL	SVM_LABEL	FORESTS_LABEL	BAGGING_LABEL	TREE_LABEL	Naïve Bayes (CARET)
2015 MMM Datasets (grouped-subset) - Machine Learning Results						
Negative	7110	9475	9809	10085	13344	12741
Neutral	21818	19217	18883	18585	16076	15509
Positive	492	728	728	750	NA	1170
2016 MMM Datasets (grouped-subset) - Machine Learning Results						
Negative	3426	4567	4741	4863	6678	5971
Neutral	11790	10482	10308	10170	8813	8831
Positive	275	442	442	458	NA	689
2016 Dover Datasets (grouped-subset) - Machine Learning Results						
Negative	1255	1440	1495	1500	1882	1849
Neutral	1876	1665	1610	1606	1292	1221
Positive	43	69	69	68	NA	104
2016 AA Datasets (grouped-subset) - Machine Learning Results						
Negative	4683	5716	5935	6034	8318	8677
Neutral	23585	22249	22030	21902	21645	18203
Positive	1695	1998	1998	2027	NA	3083
2016 Grouped-Subset with Grouped Unseen - Machine Learning Results						
Negative	16474	21198	21980	22482	30222	29238
Neutral	59069	53613	52831	52263	47826	43764
Positive	2505	3237	3237	3303	NA	5046

Table 47 MR2 Grouped (Combined) Algorithm Results for New Data

6.5.4.3 Dictionary: MR1 and MR2 Agreed Results

Both MR1 and MR2 (known as 'gold standard' refer to section 6.4.3 for further explanation) as iterated above have an agreement of over a thousand for three datasets except for Anti-Austerity which is on 840 with least agreement. To make for a fair outcome, each of the dataset will be a sample of 840, the split is 80/20, so for train is 672 and validation is 168, and then this will be tested against automated relevant datasets to predict their sentiment.

In Table 48, the NB results for each demonstration shows that negative or neutral have the highest count. In Table 48, 2015 MMM's NB classifier correctly predicted 134, where negative is 84, 50 for neutral and positive is 0. The incorrectly classified is 34 with negative on 20 and neutral on 14. The average F1 score for each sentiment class combined is 0.54, but based on single class 2015 MMM has the highest F1 score for both negative (0.84) and neutral (0.78), but 2016 MMM has the strongest positive F1 score of 0.70, which is same as for both MR1 and MR2.

CARET (R package)						
Prediction: Naïve Bayes (MMM 2015)				**Naïve Bayes (MMM 2016)**		
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
Negative	84	15	5	40	10	0
Neutral	11	50	3	27	63	6
Positive	0	0	0	4	3	15
Precision	0.81	0.78	NA	0.8	0.66	0.68
Recall	0.88	0.77	0	0.56	0.83	0.71
F1 Score	0.84	0.78	NA	0.66	0.73	0.70
Accuracy	0.80			0.70		
Prediction: Naïve Bayes (Dover)				**Naïve Bayes (Anti-Austerity)**		
Negative	146	21	1	17	5	0
Neutral	0	0	0	24	94	1
Positive	0	0	0	1	19	7
Precision	0.87	NA	NA	0.77	0.79	0.26
Recall	1	0	0	0.41	0.80	0.88
F1 Score	0.93	NA	NA	0.53	0.79	0.40
Accuracy	0.87			0.70		

Table 48 Naive Bayes results for datasets

In Table 48, for 2016 MMM there are 118 (-14 less than 2015 MMM) classified correct, with 40 (+44 more than 2015 MMM) negative, neutral is on 63 (-13) and 5 (-5) positives. The incorrectly classified is 50 with negative on 10 (+10), neutral on 33 (-19) and positive on 7 (-7). The average F1 score of each sentiment class combined is 0.70, which is 0.16 higher than

2015 MMM. NB for 2015 MMM has more correct than 2016 MMM, but 2016 has a higher F1 score as it has a greater capability to recognise the positive category. In Table 48, for Dover NB has correctly predicted 146 negative and incorrectly for neutral on 21 and positive is 1. The algorithm's F1 score is low for both neutral and positive due to the sample train data containing a higher number of negative results. This experience is the same for both MR1 and MR2 Dover. The average F1 score is 0.31, but Dover has the highest F1 score for negative (0.93) class but 0 for both neutral and positive. NB has the poorest F1 score result for Dover. In Table 48, Anti-Austerity's NB classifies 118 correctly and 44 are incorrect. The Anti-Austerity trainset contains the highest number of neutral tweets and is second place for total count of positives with 2016 MMM which is similar for both MR1 and MR2 results. The most evenly balanced dataset in each of the sentiment categories in Table 48 is 2016 MMM. The average F1 score is 0.57, but based on single sentiment classes 2015 MMM has highest F1 score for neutral is 0.79.

In Table 49, the MR1 and MR2 algorithms' F1 scores are mainly lower than MR1, but higher than MR2. The strongest algorithm is a Max Entropy of 0.66 and Forest of 0.60 with the rest lower than 0.60. Max Entropy does not have the highest correctly classified, but it is the most evenly balanced and has the most positives, which is why it has the highest F1 score. The highest correctly classified is Tree, but it has the most uneven numbers between the sentiment categories and contains no positives.

	MMM 2015					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	78	17	0	78	16	1
Neutral	9	54	2	7	57	1
Positive	3	4	1	2	4	2
Precision	0.64			0.71		
Recall	0.59			0.65		
F1 Score	0.60			0.66		
Accuracy	0.79			0.82		
	Support Vector Machine			**Bagging**		
Negative	77	18	0	77	18	0
Neutral	13	52	0	12	50	3
Positive	4	4	0	3	4	1
Precision	0.51			0.60		
Recall	0.54			0.57		
F1 Score	0.52			0.57		
Accuracy	0.77			0.76		
	Tree					
Negative	89	6	0	**Precision**	0.56	
Neutral	13	52	0	**Recall**	0.58	
Positive	5	3	0	**F1 Score**	0.57	
				Accuracy	0.84	

Table 49 2015 MMM results of other machine learning algorithms

In Table 50 the strongest algorithm for F1 is Max Entropy at 0.67 with the other algorithms all around 0.64 to 0.66. Max Entropy has the highest correctly classified of 119, but it has reasonably balanced sentiment categories and has a count of 10 positives which is the same for most algorithms except for Forest which is on 12. However, 2015 MMM Max Entropy has the second highest F1 score, as the balance of sentiment count for each category is unevenly balanced; otherwise it would have been the similar result as 2016 MMM. The worst performing algorithms are both Bagging and Tree, but are only 0.03 behind the highest F1 score.

	MMM 2016					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	48	17	6	48	21	2
Neutral	16	57	3	13	61	2
Positive	2	7	12	1	10	10
Precision	0.67			0.71		
Recall	0.67			0.65		
F1 Score	0.66			0.67		
Accuracy	0.70			0.71		
	Support Vector Machine			**Bagging**		
Negative	43	26	2	45	24	2
Neutral	12	61	3	16	58	2
Positive	3	8	10	2	9	10
Precision	0.68			0.69		
Recall	0.63			0.62		
F1 Score	0.65			0.64		
Accuracy	0.68			0.67		
	Tree					
Negative	60	9	2	**Precision**	0.68	
Neutral	32	42	2	**Recall**	0.63	
Positive	2	9	10	**F1 Score**	0.64	
				Accuracy	0.67	

Table 50 2016 MMM results of other machine learning algorithms

In Table 51 Dover's strongest algorithm for F1 is Max Entropy of 0.82 with the other algorithms ranging from 0.45 to 0.51. Max Entropy has the highest correctly classified of 152 which is in front of Tree by 1 that has a F1 score of 0.46. Max Entropy has a very high F1 score and it is the only algorithm to contain a correctly classified positive result. Dover agrees with both MMM results that Max Entropy is the best performing algorithm.

	Dover 2016					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	137	9	0	143	3	0
Neutral	12	9	0	13	8	0
Positive	1	0	0	0	0	1
Precision	0.47			0.88		
Recall	0.46			0.79		

F1 Score	0.46			0.82		
Accuracy	0.87			0.91		
	Support Vector Machine			**Bagging**		
Negative	137	9	0	135	11	0
Neutral	12	9	0	11	9	1
Positive	1	0	0	1	0	0
Precision	0.47			0.46		
Recall	0.46			0.45		
F1 Score	0.46			0.45		
Accuracy	0.87			0.86		
	Tree					
Negative	140	6	0	**Precision**	0.53	
Neutral	10	11	0	**Recall**	0.49	
Positive	1	0	0	**F1 Score**	0.51	
				Accuracy	0.90	

Table 51 2016 Dover results of other machine learning algorithms

In Table 52 Anti-Austerity's strongest algorithm for F1 is Max Entropy of 0.66 and Tree of 0.63, and the other algorithms range from 0.49 to 0.55. Additionally, the agreed Max Entropy of 0.66 is higher than the top algorithm for MR1, which is Tree on 0.65. However, these results are based on a smaller sample than both MR1 and MR2. Max Entropy has the highest correctly classified of 130 which is in front of Tree by 10 that has a F1 score of 0.60, which is why it is best performing algorithm. Anti-Austerity agrees with both MMM and Dover results that Max Entropy is the best performing algorithm. The worst performing algorithms are both Bagging and Forest on 0.49.

	Anti-Austerity 2016					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	23	19	0	24	18	0
Neutral	18	94	6	10	101	7
Positive	4	3	1	0	3	5
Precision	0.49			0.65		
Recall	0.49			0.68		
F1 Score	0.49			0.66		
Accuracy	0.70			0.77		
	Support Vector Machine			**Bagging**		
Negative	23	19	0	24	18	0
Neutral	16	98	4	19	93	6
Positive	3	3	2	4	3	1
Precision	0.57			0.49		
Recall	0.54			0.49		
F1 Score	0.55			0.49		
Accuracy	0.73			0.70		
	Tree					
Negative	26	16	0	**Precision**	0.59	
Neutral	17	87	14	**Recall**	0.75	
Positive	0	1	7	**F1 Score**	0.63	
				Accuracy	0.71	

Table 52 2016 Anti-Austerity results of other machine learning algorithms

The agreed of MR1 and MR2 results show Max Entropy is consistently the best performing dictionary except for 2016 MMM which is Naïve Bayes. The highest F1 score is 0.82 for Dover, then 2016 MMM is 0.70, and both 2015 MMM and Anti-Austerity on lowest of 0.66. Both MR1 and MR2 results show that Max Entropy is the best performing algorithm. The worst performing algorithm for both 2016 MMM and Dover is Bagging, for 2015 MMM it is SVM, and for Anti-Austerity both Forest and Tree. These results are varied on the lowest performing algorithm except for 2015 MMM with SVM worst, which agrees with MR1 and MR2. The agreed results show both 2015 MMM and Dover with highest correctly classified for negative, 2016 MMM and Anti-Austerity being neutral. The algorithm that typically has the highest level of positive correctly classified tends to have the higher F1 score, which is similar to both MR1 and MR2 results.

In the fitting of several models, the results (refer to Table 150 to Table 154 in appendix 10.12.3) are reflective of the results produced above with neutral and negative highest count, and neutral on the highest count for both 2016 MMM (exception for Tree which shows negative highest) and Anti-Austerity, and negative is the highest count for both 2015 MMM and Dover. This is re-affirmed in Table 154 in appendix 10.12.3 when all trainsets are combined of 3360 and randomly placed against the validation set of 672. Each algorithms results displays a similar proportion with negative and neutral except for Max Entropy with neutral counted 66 ahead of negative and Naïve Bayes counted 85 negative ahead of neutral. The majority of the algorithms have a dominate sentiment category count where are 4 negative and 2 neutral with positive on considerably less. Whereas, MR1 are all negative and MR2 all neutral except for Tree which had majority of negative by a few. The agreed MR1 and MR2 show a more balanced sentiment category of neutral and negative, but still not for positive which are similar to both MR1 and MR2.

The trained/validated models are now tested against the entire set of automated coded (relevant) tweets to predict their sentiment. In Table 53 the classifiers' predictions that are highest are negative and neutral, with negative dominant for Dover, but the classifiers for 2015 MMM it is 3 neutral and 3 negative, 2016 MMM 4 negative and 2 neutral, and Anti-Austerity 6 neutral. Anti-Austerity has the highest positive count by proportion than the other datasets, with Dover on zero, which is the same as both MR1 and MR2. These results are similar to MR1 and MR2 with negative and neutral being the highest, which demonstrates further to include more positive tweets in the sample to identify if it would make an impact on the overall result.

SENTIMENT_LABEL	MAXENTROPY_LABEL	SVM_LABEL	FORESTS_LABEL	BAGGING_LABEL	TREE_LABEL	Naïve Bayes (CARET)
2015 MMM Machine Learning Results: Interrater Agreement						
Negative	12639	13842	13200	13955	18156	18900
Neutral	14694	15578	14637	13530	11264	10516
Positive	2087	NA	1583	1935	NA	4
2016 MMM Machine Learning Results: Interrater Agreement						
Negative	4906	5353	6343	6071	7707	4631
Neutral	9317	8938	7994	8296	6715	8881
Positive	1268	1200	1154	1124	1069	1979
2016 DOVER Machine Learning Results: Interrater Agreement						
Negative	2678	2683	2625	2610	2628	3174
Neutral	472	491	529	535	546	NA
Positive	24	NA	20	29	NA	NA
2016 AA Machine Learning Results: Interrater Agreement						
Negative	5453	5939	6736	6627	6723	2627
Neutral	20038	21905	20398	21027	16442	18558
Positive	4472	2119	2829	2309	6798	8778

Table 53 Agreed Grouped (Combined) Algorithm Results for New Data

6.5.4.3.1 Agreed MR1 and MR2 Grouped (Combined)

As iterated in section 6.5.4.3 both MR1 and MR2 have an agreement of over a thousand for three datasets except for Anti-Austerity which is on 840 with least agreement. Each of the datasets are combined into a total of 3360, then split same 80/20, which is 2688 for train and validation is 672, and then this is tested against automated coded (relevant) datasets to predict their sentiment.

In Table 54, Naïve Bayes contains the highest F1 score of 0.70, whereas the other algorithms are all around the mid-0.60s. These agreed (both MR1 and MR2) grouped results contain higher F1 scores results rather than MR1 grouped that has the highest F1 scores over MR2 grouped. For instance, Max Entropy of 0.67, Tree of 0.68 and Naïve Bayes of 0.70 are all higher than MR1's grouped highest score of 0.66, which agreed results remaining algorithms share the same score of 0.66. Naïve Bayes has correctly classified 516 which is the second lowest, and Forest has the most correct on 526, but Naïve Bayes has the highest F1 score as it has the most correctly classified for positive. Tree is the worst performer to have correctly classified on 513, which is only 3 behind NB.

	MR1 and MR2 Grouped (Combined)					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	264	44	5	239	69	5
Neutral	60	249	5	45	266	3
Positive	9	23	13	8	23	14
Precision	0.72			0.73		
Recall	0.64			0.64		
F1 Score	0.66			0.67		
Accuracy	0.78			0.77		
	Support Vector Machine			**Bagging**		
Negative	260	49	4	258	50	5
Neutral	58	251	5	57	252	5
Positive	9	23	13	11	21	13
Precision	0.72			0.71		
Recall	0.64			0.64		
F1 Score	0.66			0.66		
Accuracy	0.78			0.78		
	Tree			**Naïve Bayes**		
Negative	254	52	7	265	77	13
Neutral	67	241	6	37	226	7
Positive	5	22	18	11	11	25
Precision	0.71			0.75	0.84	0.53
Recall	0.66			0.85	0.72	0.56
F1 Score	0.68			0.79	0.77	0.54
Accuracy	0.76			0.77		

Table 54 Grouped (combined) results of other machine learning algorithms

The trained/validated models are now tested against the entire set of automated coded (relevant) tweets to predict their sentiment. In Table 55 the dictionary machine learning results are reflective of the results in Table 154 (refer to appendix 10.12.3) that both negative and neutral have the highest count. In Table 154 (refer to appendix 10.12.3) the majority category is negative for 4 out of 6 algorithm results with neutral is not far behind, but for Table 55, neutral has the majority for 5 out of 6 algorithm results and has a far higher count and proportion for the sentiment category than Table 154 results. Naïve Bayes is the only algorithm to have its majority stay negative in both Table 55 and Table 154 (refer to appendix 10.12.3). MR1 is mainly negative and MR2 is mostly neutral, and the agreed results of MR1 and MR2 is neutral. The agreed result for each dataset is out of 840, which may mean more negatives were removed than neutral, which might be the reason why the result has a higher neutral count. The results emphasise the sentiment categories with the highest proportion e.g., neutral are likely to determine the category for new data unless it has a more even spread across each class, which is the same outcome for MR1 and MR2 results.

In section 6.5.5 is where we discuss the algorithms overall performance and identify which ones are the strongest and weakest in their predictions.

[Intentionally Left Blank]

SENTIMENT_LABEL	MAXENTROPY_LABEL	SVM_LABEL	FORESTS_LABEL	BAGGING_LABEL	TREE_LABEL	Naïve Bayes (CARET)
2015 MMM Datasets (grouped-subset) - Machine Learning Results						
Negative	12891	14294	14534	14364	14727	16338
Neutral	15205	13659	13331	13474	12903	9945
Positive	1324	1467	1555	1582	1790	3137
2016 MMM Datasets (grouped-subset) - Machine Learning Results						
Negative	6230	7225	7334	7239	7491	8142
Neutral	8514	7359	7143	7213	6931	5645
Positive	747	907	1014	1039	1069	1704
2016 Dover Datasets (grouped-subset) - Machine Learning Results						
Negative	1835	1969	2004	2018	2024	2195
Neutral	1229	1082	1033	1017	999	735
Positive	110	123	137	139	151	244
2016 AA Datasets (grouped-subset) - Machine Learning Results						
Negative	8362	9633	9690	9633	9365	12401
Neutral	18226	16719	16008	16444	16115	11345
Positive	3375	3611	4265	3886	4483	6217
2016 Grouped-Subset with Grouped Unseen - Machine Learning Results						
Negative	29318	33121	33906	33707	33607	38466
Neutral	43174	38819	37653	37852	36948	29510
Positive	5556	6108	6489	6489	7493	10072

Table 55 MR1 & MR2 Agreed Grouped Algorithm Results for New Data

6.5.5 Dictionary: Algorithm Performance Results

The dictionary machine learning performance of each algorithm will be ranked for MR1, MR2 and agreed MR1 and MR2 to help determine the best and worst algorithms based on the results. In Table 56, MR1 determines that Support Vector Machine (SVM) is the worst and Naïve Bayes is the best, with Max Entropy and Tree slightly behind, whereas remaining algorithms are much worse off.

MR1	MMM 2015	MMM 2016	Dover	Anti-Austerity	Total
Support Vector Machine	0.48	0.59	0.37	0.52	**1.96**
Bagging	0.67	0.6	0.41	0.52	**2.2**
Forest	0.63	0.6	0.42	0.51	**2.16**
Tree	0.5	0.67	0.48	0.65	**2.3**
Max Entropy	0.65	0.68	0.44	0.63	**2.4**
Naïve Bayes	0.74	0.66	0.89	0.59	**2.88**
Rank					
Support Vector Machine	6	6	6	4	**22**
Bagging	2	4	5	4	**15**
Forest	4	4	4	6	**18**
Tree	5	2	2	1	**10**
Max Entropy	3	1	3	2	**9**
Naïve Bayes	1	3	1	3	**8**

Table 56 MR1 ranked algorithm performance

In Table 57, MR2 determines that Support Vector Machine (SVM) is the worst and Naïve Bayes is the best on 5, with a Max Entropy of 8 and Forest of 12 behind it, and the remaining algorithms are lower ranked. MR2 agrees with MR1 that Naïve Bayes and Max Entropy are two of the best performing algorithms.

MR2	MMM 2015	MMM 2016	Dover	Anti-Austerity	Total
Support Vector Machine	0.42	0.36	0.38	0.3	**1.46**
Bagging	0.5	0.46	0.4	0.35	**1.71**
Forest	0.5	0.52	0.41	0.33	**1.76**
Tree	0.44	0.39	0.42	0.3	**1.55**
Max Entropy	0.53	0.53	0.47	0.33	**1.86**
Naïve Bayes	0.89	0.52	0.82	0.6	**2.83**
Rank					
Support Vector Machine	6	6	6	5	**23**
Bagging	3	4	5	2	**14**
Forest	3	2	4	3	**12**
Tree	5	5	3	5	**18**
Max Entropy	2	1	2	3	**8**
Naïve Bayes	1	2	1	1	**5**

Table 57 MR2 ranked algorithm performance

In Table 58, the MR1 and MR2 agreed determines that Bagging of 20 is worst, but Support Vector Machine (SVM) is behind on 18. The best performing algorithm is Naïve Bayes on 6, then Max Entropy of 7, Tree on 14 and Forest on 15. The agreed results agree with MR1 and MR2 that Naïve Bayes and Max Entropy are the best performing

algorithms. Additionally, agreed results agrees with MR1 that Tree is in third place except for MR2 which is Forest.

MR1 & MR2	MMM 2015	MMM 2016	Dover	Anti-Austerity	Total
Support Vector Machine	0.52	0.65	0.46	0.55	**2.18**
Bagging	0.57	0.64	0.45	0.49	**2.15**
Forest	0.6	0.66	0.46	0.49	**2.21**
Tree	0.57	0.64	0.51	0.63	**2.35**
Max Entropy	0.66	0.67	0.82	0.66	**2.81**
Naïve Bayes	0.81	0.7	0.93	0.58	**3.02**
Rank					
Support Vector Machine	6	4	4	4	18
Bagging	4	5	6	5	20
Forest	3	3	4	5	15
Tree	4	5	3	2	14
Max Entropy	2	2	2	1	7
Naïve Bayes	1	1	1	3	6

Table 58 MR1 and MR2 agreed ranked algorithm performance

In Table 59, the best performing algorithm is Naïve Bayes on 19, Max Entropy of 24 and Tree on 42 and the worst is SVM on 63. The agreed results agree with MR1 and MR2 that Naïve Bayes and Max Entropy are the best performing algorithms. These are the best performing for the dictionary approach. This will be further explored in the machine learning approach in section of 6.6 to identify if there is common agreement, as only the top three algorithms will be considered in the change point analysis.

Algorithm Category	Grand Total of Algorithm Score	Grand Total of Algorithm Rank
Support Vector Machine	5.6	63
Bagging	6.06	49
Forest	6.13	45
Tree	6.2	42
Max Entropy	7.07	24
Naïve Bayes	8.73	19

Table 59 Overall ranked algorithm performance

6.6 Machine Learning Approach: Tweets and Manual Classification

The dictionary machine learning results have concluded and now we explore the results of machine learning with only tweets and manual classification for both MR1 and the agreed MR1 and MR2 (combined) to identify if there is common agreement between each approach and which has the strongest outcome based on their F1 scores.

In section 6.6.1 the input data for train and validation will be split 80/20 from 1500 rows of tweets and the manual classification for MR1 for each dataset. However, for MR1 and MR2 agreed results in section 6.6.1.1 the trainset will group each dataset, randomise it and split it into 2688 trainset and 672 for validation to build models to predict sentiment. Additionally, the validation data will be fed against several different models (e.g., MaxEnt, SVM and Bagging) from the automated coded (relevant) data from each of the four separate datasets and a combined version of all four datasets. NB has not been used in this machine learning process as error message could not be resolved when applying the input data and instead neural network has been applied. These results will again consist of precision, recall and F1 to determine the strength of the outcome.

6.6.1 MR1 Tweets and Manual

In Table 60 2015 MMM's algorithms strongest F1 score is SVM of 0.61, which is the only algorithm in the 0.60s as the other scores are lower with Max Entropy in second on 0.55. Both SVM and Max Entropy show a higher F1 score as its classifier contains more correct positives, which agrees with dictionary approach. The worst F1 score is 0.39 for Tree.

	MMM 2015					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	127	53	1	135	38	8
Neutral	18	87	0	27	74	4
Positive	4	10	0	4	7	3
Precision	0.48			0.54		
Recall	0.51			0.55		
F1 Score	0.48			0.55		
	Support Vector Machine			**Bagging**		
Negative	138	38	5	109	70	2
Neutral	25	77	3	19	86	0
Positive	3	6	5	3	11	0
Precision	0.62			0.45		
Recall	0.62			0.47		
F1 Score	**0.61**			0.44		
	Tree			**NNETWORK**		
Negative	159	20	2	136	39	6
Neutral	71	34	0	45	53	7

Positive	11	3	0	6	5	3
Precision	0.42			0.49		
Recall	0.40			0.49		
F1 Score	0.39			0.49		

Table 60 2015 MMM results for machine learning algorithms

In Table 61 2016 MMM strongest algorithm for F1 is Max Entropy on 0.58, slightly behind is SVM of 0.57, and the remaining algorithm results are in the 0.40s. Max Entropy has correctly classified 172, but SVM is on 185. Max Entropy shows a higher F1 score as its classifier contains more correct positives, which agrees with 2015 MMM and the dictionary approach. The lowest F1 score is 0.43 for both Tree and NNetwork, which is the same for 2015 MMM.

	MMM 2016					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	58	78	4	78	55	7
Neutral	31	103	0	43	84	7
Positive	7	12	7	8	8	10
Precision	0.4766667			0.5733333		
Recall	0.50			0.5933333		
F1 Score	0.4633333			0.5766667		
	Support Vector Machine			**Bagging**		
Negative	94	43	3	59	76	5
Neutral	49	82	3	28	105	1
Positive	9	8	9	4	13	9
Precision	0.67			0.63		
Recall	0.5666667			0.5066667		
F1 Score	**0.57**			0.4833333		
	Tree			**NNETWORK**		
Negative	43	95	2	34	94	12
Neutral	16	117	1	25	89	20
Positive	2	15	9	7	4	15
Precision	0.55			0.4766667		
Recall	0.4633333			0.5133333		
F1 Score	0.4266667			0.4333333		

Table 61 2016 MMM results for machine learning algorithms

In Table 62 Dover's results strongest algorithm for F1 are Max Entropy, SVM and NNetwork on 0.38, and the remaining algorithms results are slightly lower in the 0.30s. Max Entropy has correctly classified 240, SVM is on 243 and NNetwork is 239. SVM could be deemed the strongest out of the 3 algorithms due to higher precision and recall (Max Entropy and NNetwork rounded up to 0.38), alongside with the most correctly classified, especially in the negative category.

	Dover 2016					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	238	0	0	231	6	1
Neutral	49	7	1	46	9	2

Positive	5	0	0	5	0	0
Precision	0.61			0.47		
Recall	0.37			0.3766667		
F1 Score	0.37			0.38		
	Support Vector Machine			**Bagging**		
Negative	234	3	1	236	1	1
Neutral	47	9	1	48	6	3
Positive	5	0	0	5	0	0
Precision	0.52			0.56		
Recall	0.38			0.37		
F1 Score	0.38			0.37		
	Tree			**NNETWORK**		
Negative	237	1	0	230	7	1
Neutral	51	5	1	44	9	4
Positive	5	0	0	4	1	0
Precision	0.55			0.45		
Recall	0.3633333			0.3766667		
F1 Score	0.3533333			0.38		

Table 62 2016 Dover results for machine learning algorithms

In Table 63 Anti-Austerity's strongest algorithm for F1 is SVM on 0.66 with the remaining algorithms results around the mid-0.50s except for Tree on 0.39. SVM has the most correctly classified of 221 and second is Forest on 118. SVM is the strongest out of the 2 algorithms due to higher precision, recall and accuracy, especially in the positive category.

	Anti-Austerity 2016					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	159	18	1	127	44	7
Neutral	53	56	2	37	63	11
Positive	6	2	3	2	2	7
Precision	0.66			0.54		
Recall	0.55			0.64		
F1 Score	0.58			0.57		
	Support Vector Machine			**Bagging**		
Negative	145	31	2	166	12	0
Neutral	36	69	6	63	47	1
Positive	4	0	7	8	1	2
Precision	0.6466667			0.72		
Recall	0.69			0.51		
F1 Score	0.66			0.54		
	Tree			**NNETWORK**		
Negative	169	9	0	131	44	3
Neutral	81	30	0	28	64	19
Positive	10	1	0	1	3	7
Precision	0.4666667			0.55		
Recall	0.4066667			0.65		

F1 Score	0.39	0.57

Table 63 2016 Anti-Austerity results for machine learning algorithms

In Table 64 MR1 Grouped strongest algorithms for F1 are SVM on 0.65, Forest at 0.62 and the remaining algorithm results are around the mid-0.50s except for Tree on poor score of 0.43. SVM has correctly classified 910 and Forest is on 898. SVM is strongest out of the 2 algorithms due to higher precision, recall and accuracy with the most correctly classified, especially in the positive category.

	MR1 Grouped (Combined)					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	636	57	8	550	129	22
Neutral	178	241	10	148	250	31
Positive	34	15	21	26	22	22
Precision	0.69			0.56		
Recall	0.59			0.56		
F1 Score	0.62			0.56		
	Support Vector Machine			**Bagging**		
Negative	629	66	6	658	34	9
Neutral	164	256	9	262	161	6
Positive	23	22	25	38	12	20
Precision	0.71			0.68		
Recall	0.62			0.54		
F1 Score	0.65			0.56		
	Tree			**NNETWORK**		
Negative	667	21	13	551	120	30
Neutral	371	54	4	156	212	61
Positive	49	4	17	13	27	30
Precision	0.60			0.5366667		
Recall	0.44			0.57		
F1 Score	0.43			0.5466667		

Table 64 MR1 grouped (combined) results for machine learning algorithms

SVM consistently has the strongest F1 score (mostly in 0.60s) across most data set results with Max Entropy (mainly in the 0.50s) in second except for both Anti-Austerity and MR1 Grouped is Forest. Additionally, Dover has both SVM and Max Entropy in joint first and for 2016 MMM Max Entropy is the highest by 0.01 in front of the SVM on 0.57. Anti-Austerity's (AA) strongest F1 score is SVM on 0.66 and not far behind is the MR1 Grouped SVM on 0.65, whilst the other AA results are lower 0.50s and even further lower, such as 0.38 for Dover. Overall, the algorithms with the highest f-measure are SVM and Max Entropy with Forest behind on most occasions. The worst performing algorithm is Tree which is consistent across every dataset, and the second lowest differs between Bagging and NNetwork, which are in a range from 0.30s and 0.40s except for Anti-Austerity's Bagging of 0.54 and MR1 Grouped is NNetwork on 0.55.

In the fitting of several models, the results in Table 156 to Table 159 (refer to appendix 10.13.1.1) for each dataset are reflective of the results produced in Table 65 when all trainsets are combined. MR1 and MR2 agreed results show negative as the majority for all algorithms. However, on an individual level, both MMM events shows neutral highest for Tree and Forest and the other algorithms are all neutral except for SVM on negative being 19 ahead of neutral category. The train/validation is now to be tested (refer to Table 160 to Table 163 in appendix 10.13.1.1) against the automated coded (relevant) tweets. In Table 66, the classifiers predicted sentiment categories are highest for both negative and neutral, which shares a similar result to the train/validation results in Table 65. Overall, the sample of data requires widening to include more positive tweets as this may change the result in the future.

MR1 Grouped Train Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	724	816	848	958	1087	720
Neutral	401	344	313	207	79	359
Positive	75	40	39	35	34	121

Table 65 MR1 Grouped (Combined) Model Train Results

MR1 Grouped Test Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	42777	46391	53274	65512	73992	43029
Neutral	27319	27315	21131	9546	2096	20521
Positive	7952	4342	3643	2990	1960	14768

Table 66 MR1 Grouped (Combined) Model Test Results

The MR2 tweets and manual results outlined in appendix 10.13.2 are consistently lower than MR1 except for MR2 Dover results, which contains higher F1 scores than MR1. In all the datasets results, SVM, Forest and Max Entropy has the most highest F1 scores. Furthermore, Tree and NNetwork consistently has produced the poorest F1 scores. Therefore, MR2's F1 scores agrees with MR1 on the strongest and weakest algorithms. The MR2 F1 scores are mostly in the 0.30s and 0.40s except a couple algorithms scored in the 0.50s, such as Dover's SVM on 0.52 and AA Max Entropy on 0.53. MR2's F1 results has much lower scores across each sentiment categories when compared to MR1, therefore, no further analysis required due to the poor results. The MR2 train/validation and test results reflect the number are reflective of each as the highest count lies with neutral except for Dover as negative. Both Max Entropy and NNetwork have the highest count of positive in the results, therefore, it may be the other algorithms' method are less likely to detect the positive tweets. MR2 shows a majority for neutral whereas for MR1 it is mainly negative, which occurs similarly with

MR1 and MR2 dictionary machine learning results. This may be a result of there being a difference balance between negative and neutral counts for both MR1 and MR2.

In section 6.6.1.1 MR1 and MR2 (known as 'gold standard' refer to section 6.4.3 for further explanation) agreed results will be explored in depth because of the consistently high F1 scores.

6.6.1.1 MR1 & MR2 Grouped (Combined) Tweets and Manual

As iterated in section 6.5.4.3 both MR1 and MR2 have an agreement of over a thousand for three datasets except for Anti-Austerity which is on 840 with least agreement. Each of the datasets are combined into a total of 3360, then split same 80/20, which is 2688 for train and validation is 672, and then this is tested against automated coded (relevant) datasets to predict their sentiment.

In Table 67 2015 MMM strongest algorithm for F1 is SVM on 0.71, and then Bagging of 0.66 and Forest of 0.65 which there are three other algorithms below 0.65. SVM displays a higher F1 score as its classifier contains the highest correctly classified sentiment categories and has similar count of correct positives. The poorest F1 score is 0.58 for NNetwork which has the least correctly classified.

	MMM 2015					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	68	27	0	79	14	2
Neutral	4	60	1	14	47	4
Positive	2	4	2	3	2	3
Precision	0.75			0.63		
Recall	0.63			0.64		
F1 Score	0.65			0.63		
	Support Vector Machine			**Bagging**		
Negative	83	11	1	61	34	0
Neutral	11	53	1	6	58	1
Positive	3	2	3	1	4	3
Precision	0.75			0.75		
Recall	0.69			0.64		
F1 Score	0.71			0.66		
	Tree			**NNETWORK**		
Negative	55	40	0	55	36	4
Neutral	6	59	0	7	55	3
Positive	2	4	2	2	3	3
Precision	0.81			0.58		
Recall	0.58			0.60		
F1 Score	0.60			0.58		

Table 67 MR1 & MR2 2015 MMM results for machine learning algorithms

In Table 68 2016 MMM strongest algorithm for F1 is SVM on 0.69 and not far behind are Max Entropy on 0.67 and Forest on 0.62. The other algorithms are equal or lower than 0.60, such as Tree on 0.60 and Bagging on 0.58. SVM displays a higher F1 score as its classifier contains the highest correctly classified of 114 and has one of the highest counts of correct positives. The lowest F1 score is 0.49 for NNetwork which has the least correctly classified on 93.

	MMM 2016					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	37	32	2	45	24	2
Neutral	8	68	0	19	53	4
Positive	3	10	8	2	5	14
Precision	0.73			0.68		
Recall	0.5966667			0.67		
F1 Score	0.62			0.67		
Accuracy	0.67			0.67		
	Support Vector Machine			**Bagging**		
Negative	43	26	2	37	32	2
Neutral	17	57	2	14	62	0
Positive	1	6	14	5	9	7
Precision	0.7066667			0.68		
Recall	0.6766667			0.56		
F1 Score	**0.6866667**			0.58		
Accuracy	0.6785714			0.63		
	Tree			**NNETWORK**		
Negative	40	29	2	31	20	20
Neutral	12	64	0	10	54	12
Positive	4	10	7	9	4	8
Precision	0.7033333			0.50		
Recall	0.5766667			0.51		
F1 Score	0.60			0.49		
Accuracy	0.66			0.55		

Table 68 MR1 & MR2 2016 MMM results for machine learning algorithms

In Table 69 Dover's strongest algorithm is SVM of 0.58, then NNetwork of 0.50 and Forest/Max Entropy of 0.48 with the remaining algorithms in the lower 0.40s with Bagging on the lowest score of 0.42. SVM displays a higher F1 score as its classifier contains the highest correctly classified sentiment categories of 157 and no algorithms contain positives. The poorest F1 score is 0.42 for Bagging which has the least correctly classified on 146.

	Dover 2016					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	145	1	0	136	8	2
Neutral	14	7	0	11	10	0
Positive	1	0	0	1	0	0
Precision	0.60			0.49		
Recall	0.44			0.47		

F1 Score	0.48			0.48		
	Support Vector Machine			**Bagging**		
Negative	143	3	0	141	4	1
Neutral	7	14	0	15	5	1
Positive	1	0	0	1	0	0
Precision	0.59			0.49		
Recall	0.55			0.40		
F1 Score	**0.57**			0.42		
	Tree			**NNETWORK**		
Negative	140	6	0	136	9	1
Neutral	13	8	0	8	12	1
Positive	1	0	0	0	1	0
Precision	0.4933333			0.50		
Recall	0.4466667			0.50		
F1 Score	0.4633333			0.50		

Table 69 MR1 & MR2 2016 Dover results for machine learning algorithms

In Table 70 Anti-Austerity's strongest algorithm for F1 score is SVM of 0.69, and then follows Max Entropy of 0.67 and Forest of 0.66 with the other remaining algorithms in the lower 0.50s or 0.40s with the lowest NNetwork on 0.43. SVM displays a higher F1 score as its classifier contains the 2nd highest correctly classified sentiment categories of 130 and joint highest on positive count. The highest correctly classified count is Forest on 133, but it has one less in the positive, which is why it has a lower F1 score. The poorest F1 score is 0.43 for NNetwork which has the least correctly classified on 111.

	Anti-Austerity 2016					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	15	27	0	25	17	0
Neutral	3	114	1	24	92	2
Positive	0	4	4	1	2	5
Precision	0.81			0.68		
Recall	0.61			0.6666667		
F1 Score	0.66			0.67		
	Support Vector Machine			**Bagging**		
Negative	23	19	0	13	29	0
Neutral	14	102	2	3	112	3
Positive	1	2	5	0	5	3
Precision	0.7166667			0.6933333		
Recall	0.6766667			0.5466667		
F1 Score	0.69			0.58		
	Tree			**NNETWORK**		
Negative	13	29	0	11	21	10
Neutral	2	114	2	7	98	13
Positive	0	5	3	3	3	2
Precision	0.7466667			0.4666667		
Recall	0.5533333			0.4466667		
F1 Score	0.60			0.43		

Table 70 MR1 & MR2 2016 Anti-Austerity results for machine learning algorithms

In Table 71, MR1 & MR2 Grouped strongest algorithm for F1 score is SVM of 0.74, then Max Entropy of 0.72 and Forest of 0.65 with remaining algorithms in either lower 0.60s or 0.50s and the lowest is Tree on 0.48. SVM displays a higher F1 score as its classifier contains the highest correctly classified sentiment categories of 554 and is 4th highest on positive count. The poorest F1 score is 0.48 for Tree which has the least correctly classified on 406. MR1 and MR2 Grouped F1 has shown a significant lift in Max Entropy and SVM in 0.70s compared with the individual results above which have mainly been in 0.60s. Forest remains in the 0.60s on most individual results. The remaining algorithms are similarly lower to the individual results.

	MR1 & MR2 Grouped (Combined)					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	222	90	5	250	64	3
Neutral	27	289	1	68	234	15
Positive	9	19	10	6	6	26
Precision	0.7366667			0.71		
Recall	0.6233333			0.7366667		
F1 Score	0.65			0.72		
	Support Vector Machine			**Bagging**		
Negative	276	37	4	167	144	6
Neutral	49	259	9	16	299	2
Positive	9	10	19	4	22	12
Precision	0.7566667			0.71		
Recall	0.73			0.5966667		
F1 Score	0.74			0.61		
	Tree			**NNETWORK**		
Negative	99	209	9	156	127	34
Neutral	17	299	1	49	247	21
Positive	4	26	8	8	5	25
Precision	0.6066667			0.5633333		
Recall	0.4866667			0.6433333		
F1 Score	0.4766667			0.5733333		

Table 71 MR1 & MR2 Grouped (Combined) Machine Learning Results

SVM consistently has the highest F1 score (mostly in late 0.60s or early 0.70s) across most data set results with Max Entropy (mainly in the mid-0.60s) in second except for 2015 MMM it is Bagging 0.66 and Dover is NNetwork 0.50. The highest F1 score for individual set is SVM of 0.71 for 2015 MMM, but the grouped set is higher on 0.74. The other individual algorithm results, besides SVM, are mostly either a little lower or considerably lower in the 0.40s with Dover's Bagging on the lowest F1 score of 0.42. The best performing algorithms are mainly both SVM and Max Entropy, whilst the worst performing is mostly NNetwork which is consistent across every dataset except for Dover it is Bagging 0.42 and MR1 and MR2 grouped is Tree 0.48 with NNetwork on 0.57 in second lowest.

In the fitting of several models, the train results (refer to Table 179 to Table 182 in appendix 10.13.3) are reflective of the results produced above, and re-affirmed in Table 72 train results when MR1 and MR2 are combined into 2688 trainset and a validation of 672. These numbers show a similar split by proportion when compared to the first four standalone datasets (refer to Table 179 to Table 182 in appendix 10.13.3), showing neutral in most cases except for Dover which contains mostly negative data in train, hence the negative result. Additionally, both SVM and Max Entropy have a higher negative count in 2015 MMM and MR1 and MR2 combined.

In Table 73 the MR1 and MR2 combined trainset/validation can be tested (refer to Table 183 to Table 186 in appendix 10.13.3) against the automated coded (relevant) tweets. In Table 73 MR1 and MR2 combined are tested again all the datasets automated (coded) relevant tweets that has predicted the highest count for most classifiers are neutral, then followed by negative and positive, which is similar to the train results. The individual datasets tested against MR1 and MR2 combined is that 2015 MMM shows Max Entropy with a higher negative count over neutral and positive which is similar to the individual 2015 MMM train results. Overall, the sample of data requires widening to include more positive tweets as this may change the result in the future.

MR1 & MR2 Grouped (Combined) Train Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	324	334	258	187	120	213
Neutral	304	306	398	465	534	379
Positive	44	32	16	20	18	80

Table 72 MR1 & MR2 Grouped (Combined) Model Train Results

MR1 & MR2 Grouped (Combined) Test Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	32251	29927	17437	4648	4648	20885
Neutral	37778	42999	57106	71346	71346	42429
Positive	8019	5122	3505	2054	2054	14734

Table 73 MR1 & MR2 Grouped (Combined) Test Results

The MR1 and MR2 agreed (separate/combined) results for most algorithms are at a consistent range for the F1 scores rather than MR1 results where some algorithms can be higher or lower in range more often. When MR1 and MR2 agreed groups the individual trainsets and validation takes place the F1 score increases to a higher level compared to the separate trainset/validation results. In most of the results above shows SVM and Max Entropy are mostly highest F1 scores. Furthermore, NNetwork

consistently has the poorest F1 scores. Both MR1 and MR2 agreed (separate/combined) results are similar with MR1 and MR2 on which are best and worst performing F1 scores. MR1 and MR2 grouped (combined) F1 scores are mainly in the 0.60s and early 0.70s, which is greater than MR1 that is in both 0.60s and 0.50s. MR1 and MR2 grouped (combined) train and test results reflect the algorithms classified sentiment count, as the highest category is neutral with negative and positive behind except for Dover on negative. MR1 and MR2 grouped (combined) results agrees similarly that neutral is the highest sentiment category same as MR2, but disagrees with MR1 on the majority of negative. MR1 and MR2 agreed (separate/combined) contains less manually classified negative because of the proportion of disagreement between MR1 and MR2, hence neutral has the highest count. MR1 and MR2 grouped (combined) agrees with MR2 that both Max Entropy and NNetwork have the highest positive detection rate.

6.6.1.2 Machine Learning Approach: Tweets and Manual Algorithm Performance Results

The machine learning performance of each algorithm will be ranked for MR1, MR2 and agreed MR1 and MR2 to help determine the best and worst algorithms based on the results. In Table 74, MR1 determines that Support Vector Machine (SVM) is the best and Tree is the worst, with Max Entropy slightly behind, whereas remaining algorithms are in worse positions. SVM has gone from one of the worst to best, but the constant strongest algorithms for both dictionary and ML algorithm is Max Entropy 2nd and Forest 3rd position.

MR1	MMM 2015	MMM 2016	Dover	Anti-Austerity	Total
Support Vector Machine	0.61	0.57	0.38	0.66	2.22
Bagging	0.44	0.48	0.37	0.54	1.83
Forest	0.48	0.46	0.37	0.58	1.89
Tree	0.39	0.43	0.35	0.39	1.56
Max Entropy	0.55	0.58	0.38	0.57	2.08
NNetwork	0.49	0.43	0.38	0.57	1.87
Rank					
Support Vector Machine	1	2	1	1	5
Bagging	5	3	4	5	17
Forest	4	4	4	2	14
Tree	6	5	6	6	23
Max Entropy	2	1	1	3	7
NNetwork	3	5	1	3	12

Table 74 MR1 ranked algorithm performance for ML approach

In Table 75, MR2 determines that Max Entropy is the best on 6 and Tree is the worst on 19, with SVM of 8 and Forest of 14 behind them, and the remaining algorithms are lower ranked. MR2 agrees with MR1 that Max Entropy and SVM are two of the best

performing algorithms and Tree is the worst. SVM went from worst to best, and the constant are max entropy in 2nd and Forest in 3rd.

MR2	MMM 2015	MMM 2016	Dover	Anti-Austerity	Total
Support Vector Machine	0.48	0.35	0.52	0.44	1.79
Bagging	0.46	0.33	0.32	0.46	1.57
Forest	0.46	0.32	0.36	0.47	1.61
Tree	0.42	0.35	0.29	0.42	1.48
Max Entropy	0.46	0.37	0.45	0.53	1.81
NNetwork	0.4	0.34	0.41	0.43	1.58
Rank					
Support Vector Machine	1	2	1	4	8
Bagging	2	5	5	3	15
Forest	2	6	4	2	14
Tree	5	2	6	6	19
Max Entropy	2	1	2	1	6
NNetwork	6	4	3	5	18

Table 75 MR2 ranked algorithm performance for ML approach

In Table 76, MR1 and MR2 agreed determines that NNetwork is the worst on 20, but both Bagging and Tree are second worst on 18, whereas the dictionary approach outlined Bagging to be the worst, which is not far in agreement with the ML approach. The major difference is SVM is ranked first, then being second bottom for dictionary approach. The best performing algorithm is SVM on 4, then Max Entropy of 11, and Forest of 12. The agreed results agree with MR1 and MR2 that SVM and Max Entropy are the best performing algorithms. Additionally, the grouped results agree with MR2 that Forest is in third place except for MR1 which is NNetwork.

MR1 & MR2 grouped	MMM 2015	MMM 2016	Dover	Anti-Austerity	Total
Support Vector Machine	0.71	0.69	0.57	0.69	2.66
Bagging	0.66	0.58	0.42	0.58	2.24
Forest	0.65	0.62	0.48	0.66	2.41
Tree	0.6	0.6	0.46	0.6	2.26
Max Entropy	0.63	0.67	0.48	0.67	2.45
NNetwork	0.58	0.49	0.5	0.43	2
Rank					
Support Vector Machine	1	1	1	1	4
Bagging	2	5	6	5	18
Forest	3	3	3	3	12
Tree	5	4	5	4	18
Max Entropy	4	2	3	2	11
NNetwork	6	6	2	6	20

Table 76 MR1 & MR2 grouped (combined) - ranked algorithm performance for ML approach

In Table 77, the best performing algorithm is SVM on 17, and then Max Entropy of 24 (same as dictionary approach), Forest (Tree for Dictionary Approach) on 40 and the worst is Tree (not the same as it was SVM for dictionary approach) on 60. The grouped results agree with MR1 and MR2 that SVM and Max Entropy are the best performing algorithms. ML approach agrees with Dictionary Approach that Max Entropy is one of

the best performing algorithms, however, there is disagreement on SVM, as ML has it in the highest position, but it is one of the worst for the dictionary approach.

Algorithm Category	Grand Total of Algorithm Score	Grand Total of Algorithm Rank
Tree	5.3	60
Bagging	5.64	50
NNetwork	5.45	50
Forest	5.91	40
Max Entropy	6.34	24
Support Vector Machine	6.67	17

Table 77 Overall ranked algorithm performance for ML approach

The results from the dictionary approach overall have a stronger outcome compared to the machine learning approach based on F1 score.

The algorithms ranked in the grand total for both approaches agree Max Entropy is the one to keep on using in change point, but the continuous use of other algorithms is somewhat less clear. Based on the fact dictionary approach is in a stronger position, Naïve Bayes will be applied alongside Max Entropy in the change point analysis in section 6.7.

6.7 Change point results

The change points results are based on the manually coded (relevant) tweets and the tweets that are automatically coded as relevant with the keywords list created.

The first part of the analysis will explore the initial dictionaries sentiment-based results over time for both manually and automated coded (relevant) tweets. For the second part, the graphs are based on tweets are counted for each day and by the hour, which is divided by the total of negative, neutral and positive to produce percentage of the proportion over time. Moreover, the count is based on the average score for every dictionary individual tweet, which is then categorised into negative, positive and neutral. In the third part, several changepoint techniques were explored, such as BinSeg, PELT, AMOC, SegNeigh, and decided to go with the first choice of BinSeg, as there was no difference with other methods or even its configuration tweaked with different penalties, minseglen or cpttype (Killick, 2016). BinSeg known as Binary Segmentation method is used for identifying changepoints which provided a set of summary statistics for a specified cost function and penalty, which identifies the maximum number of changepoints to search for the timeline of events (Killick, 2016).

The graphs description for negative and neutral aligns more with the bulleted list for the timeline of events, but for positive this aligns more highly with the tweets, as most publications about the events focus on being neutral and/or negative about the event. Furthermore, the bulleted timeline of the reported events for all datasets provides

more or less information in the publications and also at specific periods of time throughout the event, which is a further reason why a description of the graphs are reliant at times on the tweets discussion rather than the reports in the bulleted list.

6.7.1 MMM 2015

Both Figure 6.1 and Figure 6.2, the results of the majority vote of the sentiment categories are shown by day and hour over a six-day period. There is a total count of 3,296 tweets for manual and 29,420 for automated. These graphs show a similar pattern to section 5.12.1.1, where it builds the day before and on the day of the event which reaches its highest peak. The sentiment for 2015 MMM is mainly negative, closely followed by neutral with a smaller number of positive.

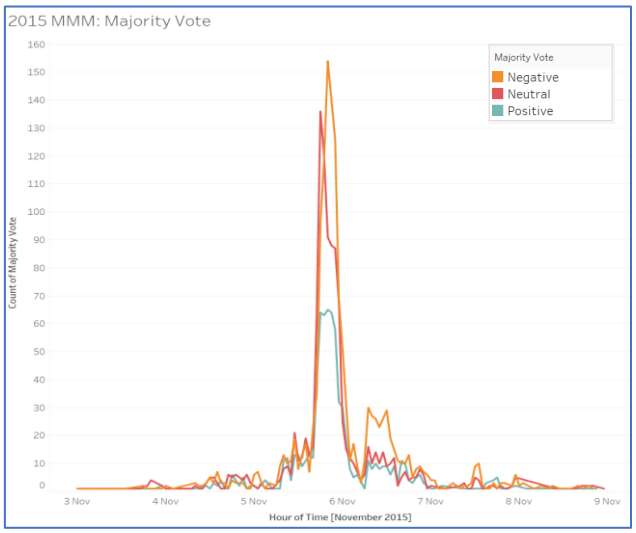

Figure 6.1 2015 MMM sentiment by day/hour (manual)

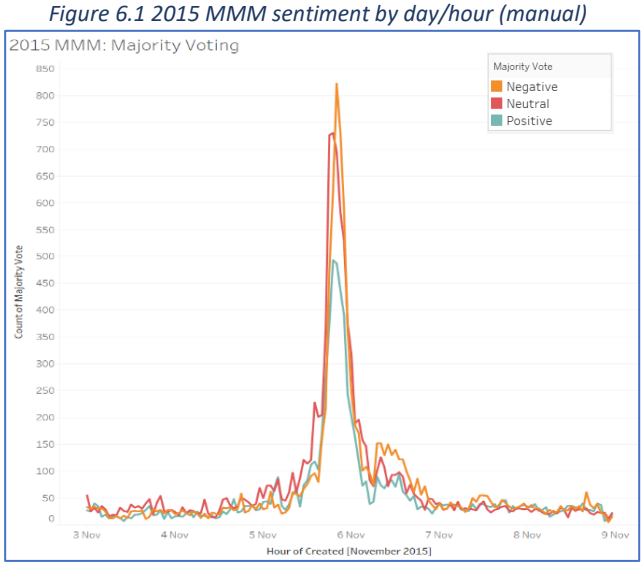

Figure 6.2 2015 MMM sentiment by day/hour (automated)

During the 2015 MMM certain events are known to have happened and these have been labelled on Figure 6.3 and Figure 6.4, but below provides a detailed description of the events and they can be seen to coincide with some of the peaks and troughs in the sentiment (Turner & Finnigan, 2015; Gayle & Johnston, 2015a & 2015b): -

- Fireworks are let off near parliament at 18:46. Fireworks are aimed at police horses in parliament square at 18:52.
- Riot police arrive at 19:00 that help hold back the demonstrators by 19:35. A police car is set on fire at 19:54 and flares let off by Big Ben at 20:15.
- Crowds are gathering at Buckingham Palace outside the permitted route at 20:23. Three men are reportedly arrested in Trafalgar Square by 20:40.
- A photographer hit by Aston Martin at 21:07, near Victoria. There are more reports of injury and violence at 21:10, and a video of kettling by police at Parliament Square at 21:17. Demonstrators carry a coffin full of money through the streets.
- Police attempt to disperse the crowds at 22:20 and the end time of demonstration is 21:00. Police warn lingering demonstrators at 22:25 could be arrested. Crowds remain in Trafalgar Square at 22:39 and police set up two containments in central London to disperse remaining demonstrators at 23:30.

Both Figure 6.3 and Figure 6.4 show the percentage share of neutral, negative and positive tweets throughout the key time period leading up to, during and after the event (between 16:00 to midnight). The number of tweets is depicted by the thickness of each trace. In Figure 6.3 we are now analysing a specific time block of which the neutral category displays the highest percentage of tweets throughout the event and peaks at 64% at 18:00, but the 'Tweets Count' shows a larger line from 18:00 to 23:00. The negative category tends to be over a 33% negative and positive mostly below 20%. The neutral tweets comprised mainly factual statements and can be seen to be particularly high at the start of the event and then showed a steady decline in volume. The negative strand highest peak is at 39% and included tweets that were sarcastic about the event, later there is a peak about the affect of a police car on fire and a photographer hit by a car. Negative showed a similar decline as did neutral initially, however, 17:00 saw a major drop in the volume of tweets before it rose again at the start time of the event at 18:00, which is unusual, this is likely due to tweets left out in the manually coded dataset. The positive line is low (under 20%) and shows minor rise and fall throughout. These tweets included showing respect to police and demonstrators and outlining the event a success.

[Intentionally Left Blank]

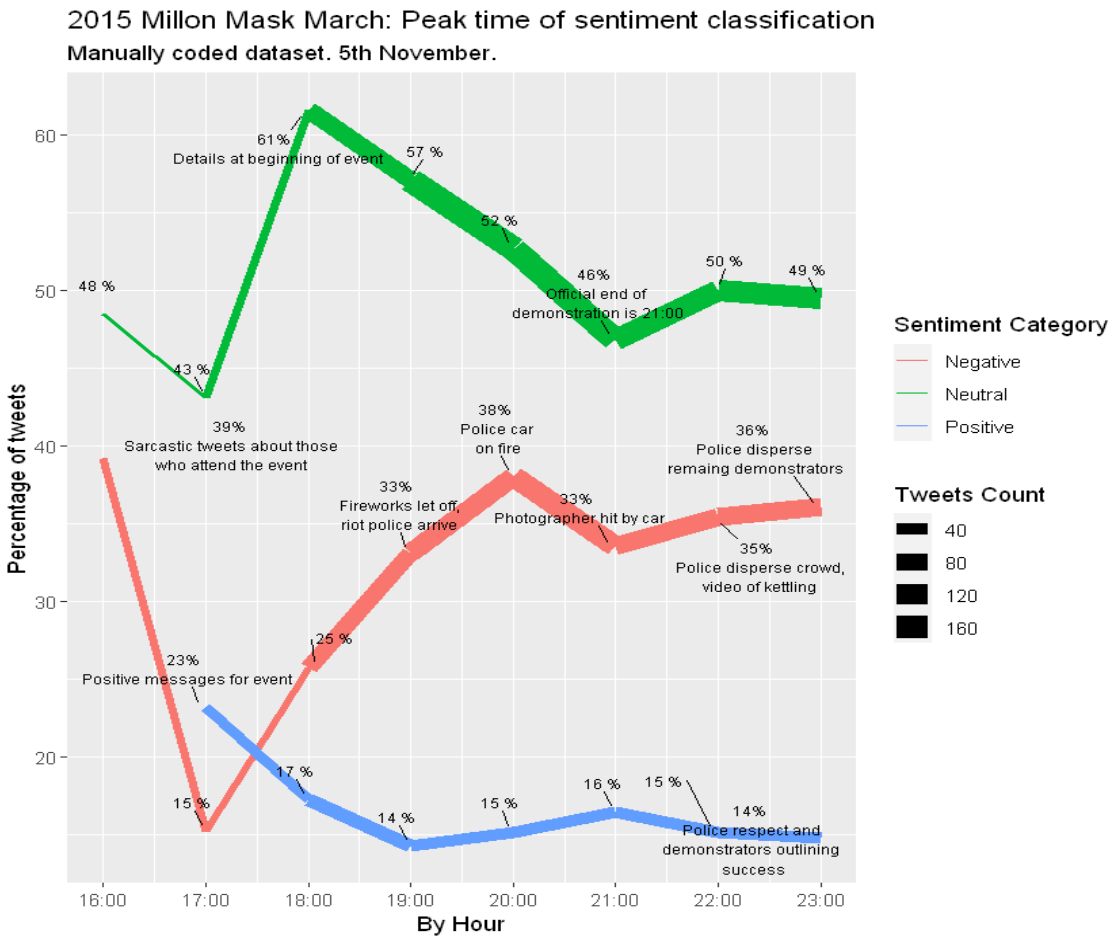

Figure 6.3 MMM 2015: Peak time of sentiment classification (manual)

In Figure 6.4, shows the neutral and negative are intertwined from 19:00 to 23:00, and seem to mirror each other, which given they are percentages and the fairly level positive share is to be expected. The tweets' count are far greater in volume than Figure 6.3, which is the reason for the dramatic change, however, the information outlined at the peaks and troughs are similar to Figure 6.3 despite the sentiment category being apart. The neutral line is highest most of the time except at 20:00 when negative peaks above neutral but positive displays similar results to Figure 6.3.

[Intentionally Left Blank]

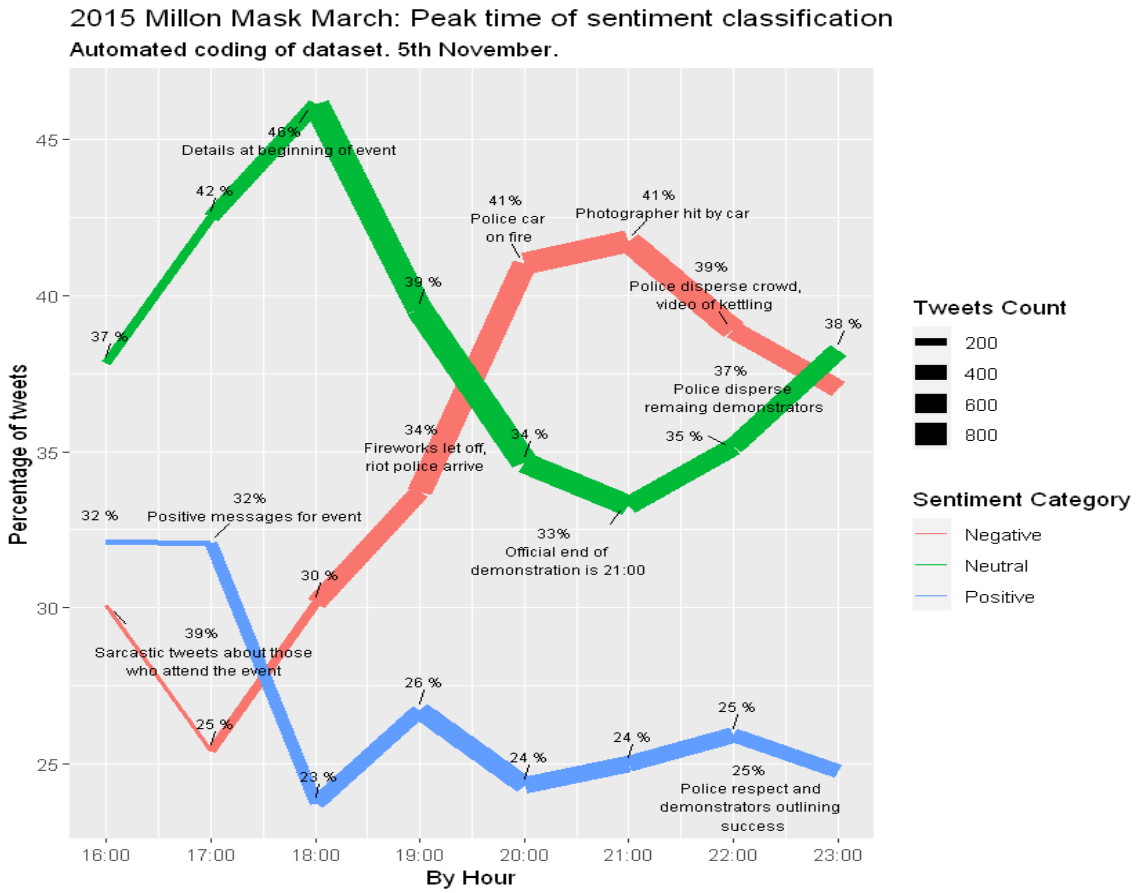

Figure 6.4 MMM 2015: Peak time of sentiment classification (automated)

In Figure 6.5, shows another way of illustrating the sentiment. The red dots are individual tweet sentiment score and the line the average of tweets for a particular time slot on the day of the event. The high volume of red dots (individual tweets) is between 18:00 and 23:00. There was higher rise in tweets from 17:00 of 138 onwards to 18:00 when the event begun, 18:00 to 19:00 is 296 which sees a surge in tweets between 19:00 and 22:00. The peak was between 20:00 and 21:00 of 310 tweets. The number of tweets is steady until after 23:00, which decreases by 102 tweets. Overall, the average sentiment remained pretty level throughout Figure 6.5.

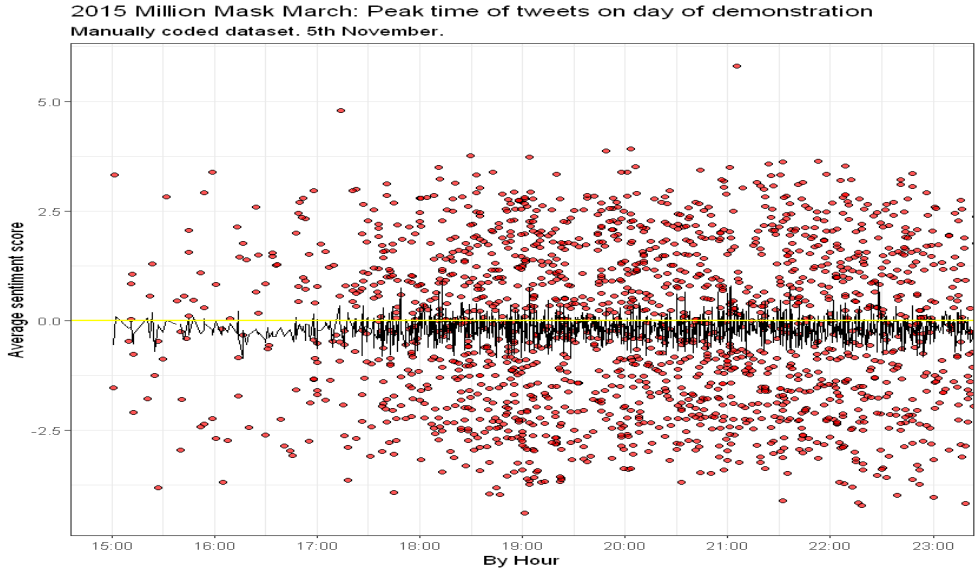

Figure 6.5 2015 MMM - Peak time of tweets on day of demonstration (Manual)

Figure 6.6 is similar to Figure 6.5, albeit with more tweets from the automated coded (relevant) datasets, but specifically this focuses on 2015 MMM. Figure 6.6 average seems to be the nearest hour and the scale on the horizontal axis only goes from 0.1 to -0.2.

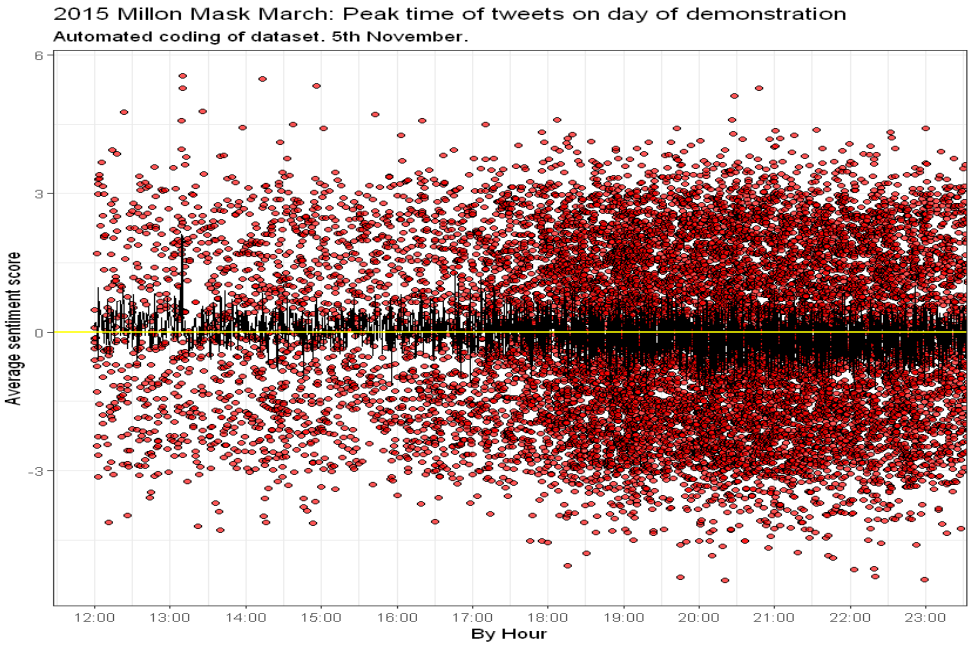

Figure 6.6 2015 MMM - Peak time of tweets on day of demonstration (Automated)

[Intentionally Left Blank]

Figure 6.7 provides a detailed view of the average score for the manually coded (relevant) tweets between 15:00 and midnight, which mimics Figure 6.5 (for the manually coded (relevant) tweets). As a result, the score is mainly on the negative side except for at 15:00 and 17:30 which shows positive tweets about the event between 0.07 to 0.10. At 19:00 it is -0.17 onwards and the negativity grew over time to a peak of -0.23 at 22:00. However, 21:00 saw less negativity, but then it rose towards 22:00. At this point a police containment arrived to disperse the demonstrators. It really only departs a less similar path towards the latter hours as negativity decreased towards 22:00 coinciding with when a police containment was sent to disperse the demonstrators.

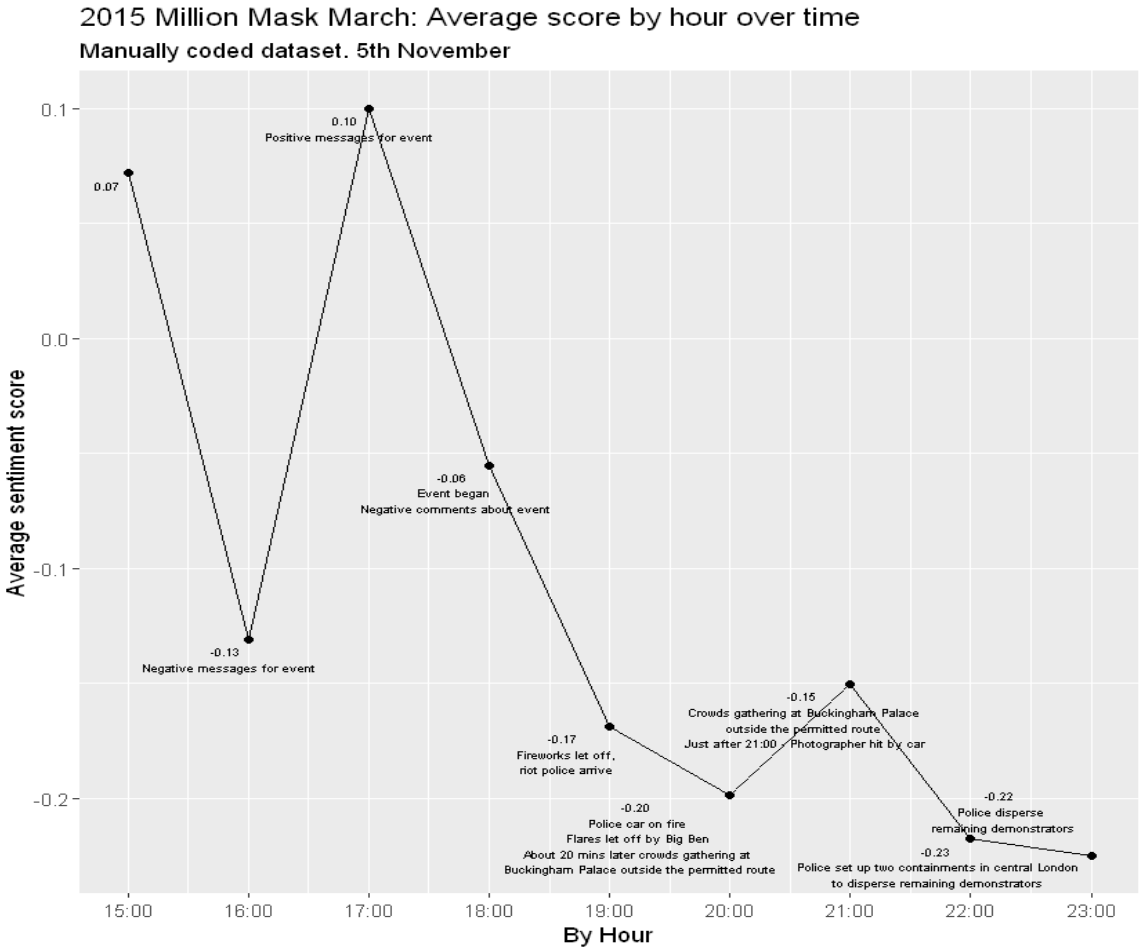

Figure 6.7 2015 MMM - Average score by hour overtime (manual)

Figure 6.8 is a detailed view of the average score for the automated coded (relevant) tweets, which has the highest negative score of -0.17 compared to -0.23 for Figure 6.7. This is due to the higher volume of tweets having more positivity. In Figure 6.8, is a detailed view of the average score for the automated coded tweets same as Figure 6.7 manually coded tweets. Similar to the above graph, the score is mainly on the negative side, agrees with Figure 6.7 that between 15:00 and 17:30 shows positive tweets about the event between 0.05 to 0.07. At 19:00 is -0.15 onwards the negativity grew over time to a peak at 21:00 of -0.17, but saw less negativity 21:00, but follows less similar path to Figure 6.7 as negativity decreased towards 22:00, at this point a police

containment to disperse the demonstrators. Figure 6.8 highest negative score of -0.17 shows less negativity than Figure 6.7 score of -0.23, this is due to the higher volume of tweets having a more positivity.

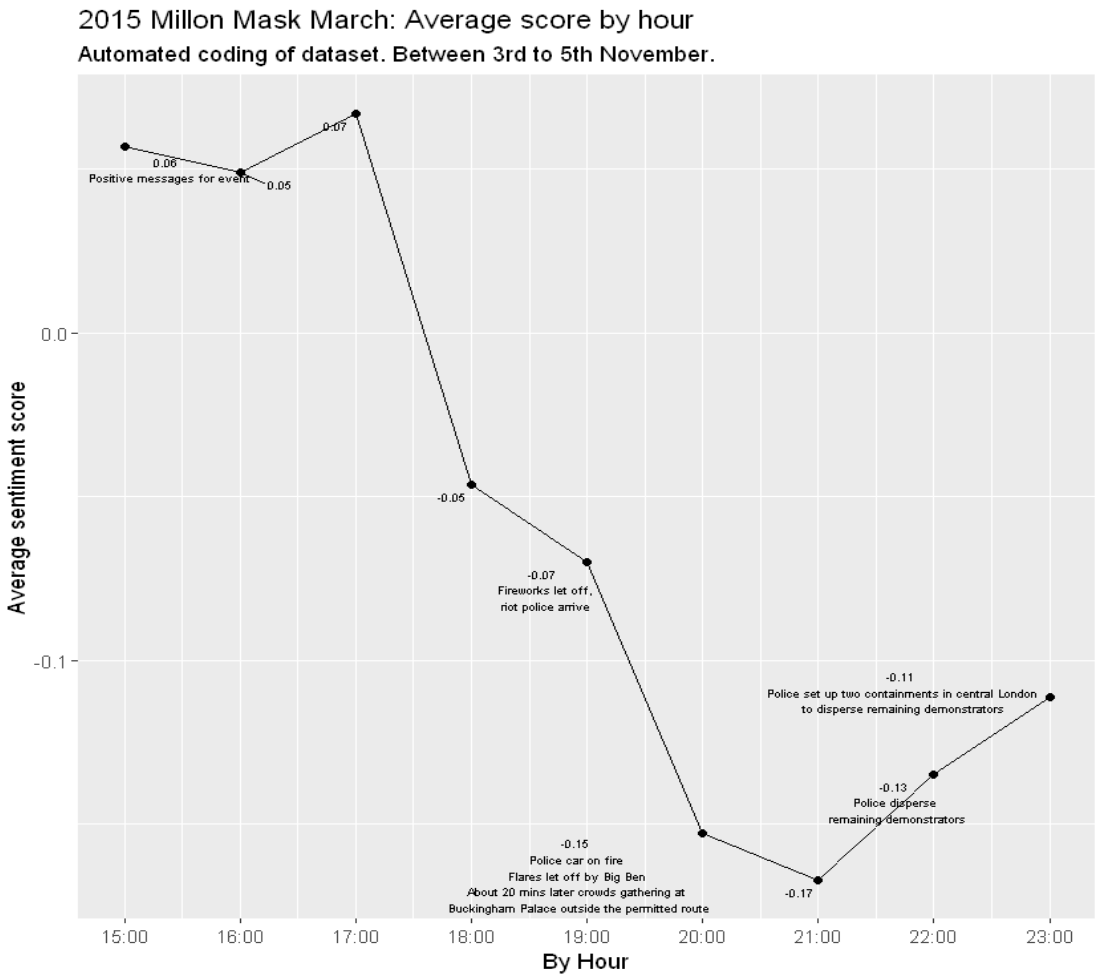

Figure 6.8 2015 MMM - Average score by hour overtime (automated)

We have explored the sentiment categories volume over time, average score and volume of tweets, which can help to determine the significant occurrences over time. In Figure 6.9, we have applied Binary Segmentation (BinSeg) to identify changepoints for negative, neutral and positive categories in a set period of time that had more than 10 tweets for the manually coded set as the sample is of a small size. The red line in the graph is the maximum number of changepoints to search for in the data, and the limit is 8 as it would mark too many points which would render it less meaningful.

In Figure 6.9, the reduction of the sample set to not include less than 10 tweets has impacted the timeline as there are gaps in the time. However, in Figure 6.9 for negative we can identify that the first two red dotted change points correspond to the time leading up to the event and the information circulated about it, together with the details and opinions surrounding this in response. The third change point is at 19:00 and corresponds to when fireworks were being let off and the riot police were making their arrival, which these turns of events caused this 'spark' in tension. On 6th of

November a change point occurred at 03:00 where the conversation is negative towards the government, police and religion with offensive language used, and at 15:00 that day has discussion around the outcome of the demonstration, which leant towards the negative impact of the demonstration and MMM reputation.

In Figure 6.9 the neutral line is made of red dotted change points where statements are made, for instance, at 14:00 05/11 the tweets suggested the importance of demonstration as petitions are not as helpful, respect the police and for police to protect the city, and other points on that day are reticent of what is previously outlined, it further supports what has gone before. The positive line has less volume of tweets for positivity which is why there are many gaps in the timeline. At 18:00 on 05/11 the tweets suggest it is about good spirits, solidarity and support for a peaceful demonstration, and at 15:00 06/11 it is thanking the demonstrator raising flag at London rally, thanks to police officers keep public safe and beautiful scenes at the event.

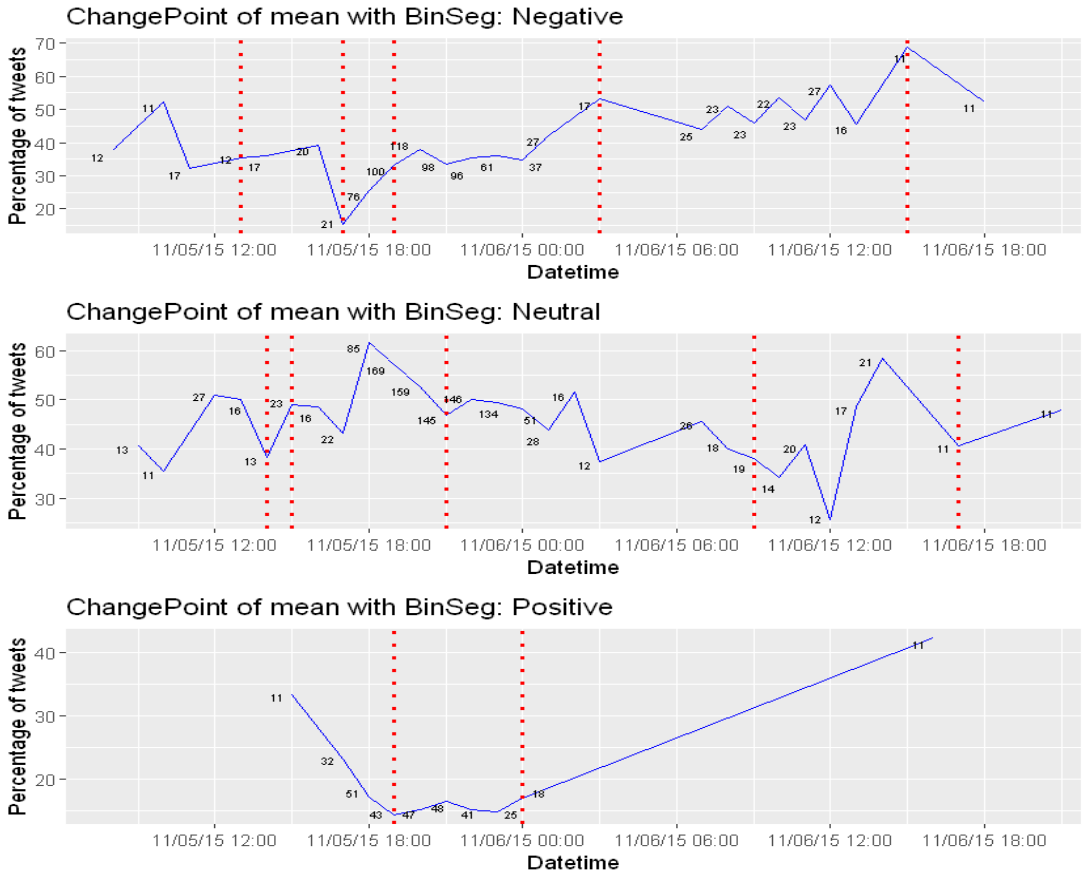

Figure 6.9 ChangePoint of mean with BinSeg by sentiment classification (manual)

Figure 6.10 shows a slightly different pattern to Figure 6.9 due the larger volume of tweets in this sample, but the overall trace does mimic the smaller sample. Based on time during the event, the negative line in Figure 6.10 has identified change point that is stronger at 19:00 to 20:00 which matches the reported reasons of a police car set on fire, flares let off and crowds outside the permitted route and, at 23:00 the next change point appears to link with reported news at 23:30, where police set up

containments to disperse remaining demonstrators. On the neutral line before the event, there are two change points between 3:00 and 4:00, which both are discussed by Twitter users "being in this together, to remember the 5th of November, and peaceful protest and use of masks". These two change points does not make sense as there is much lesser volume of tweets in that range of time, and after exploring the R documentation for BinSeg we are not sure why this has been identified. At 20:00 there is a change point same as the negative line, which the tweets outline information on "live streaming, where the news coverage is, at London, remembering Guy Fawkes and about being in it together/join revolution". After the event, Twitter users are sharing information about "costumes, heavy police presence and participation along with some that appear to share video or images".

In Figure 6.10 the positive line only displays change points after the event on the 6th of November. This may have been due to the MMM being a global event, where some data may relate to another country in a different time zone, such as the US which does appear unrelated to UK event and is not reported in UK news. For instance, the change point at 05:00 is discussion on Twitter about "peaceful protest, follow us, event goes well, cool pictures and seek truth". The changepoint at 08:00, questions on Twitter around whether a "minority or wider support for MMM, our children deserving better, and some are sarcasm and are negative about event". The change points at 20:00 to 21:00 the tweets describe the "love for the event, stand up for world of love, favourite supporter of activist group, respect and peaceful protest".

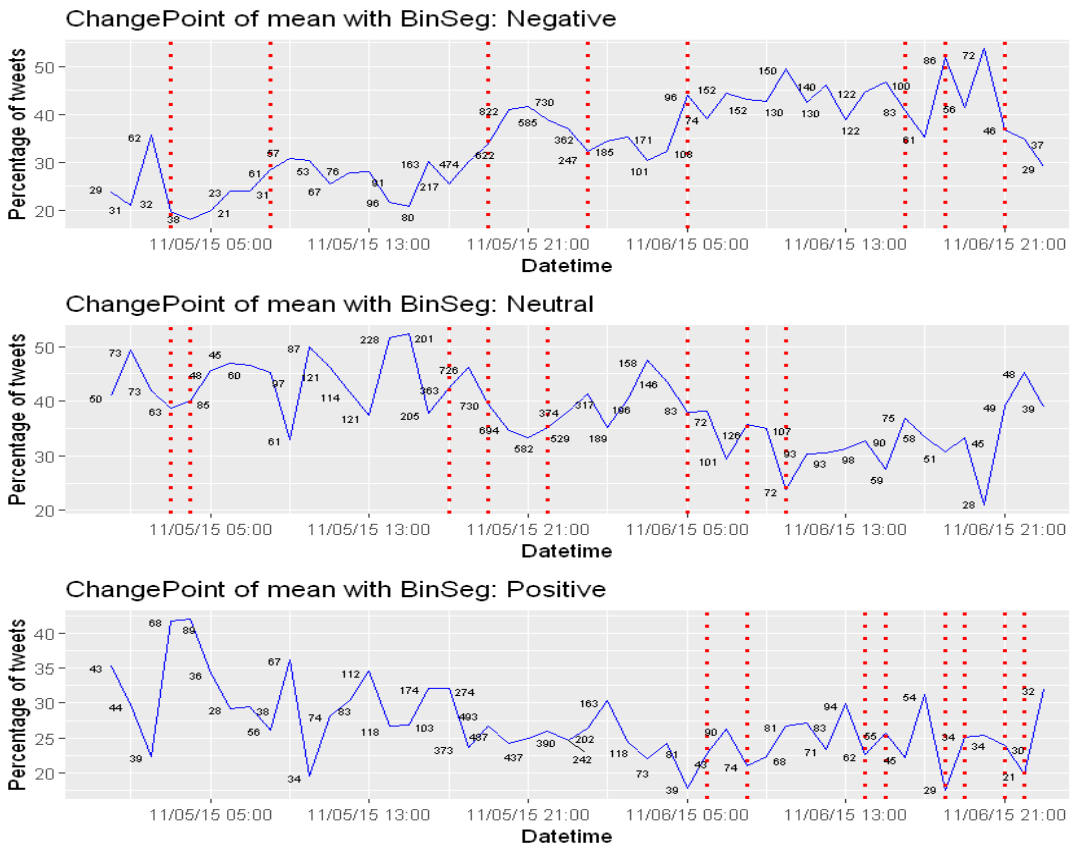

Figure 6.10 ChangePoint of mean with BinSeg by sentiment classification (automated)

We can repeat the analyses for the other techniques used to score the sentiment. As an illustration the machine learning results of the predictions from both Naïve Bayes (NB) and Max Entropy (MaxEnt) are below. In Figure 6.11, in Naïve Bayes displays a majority for negative, then neutral category with positive non-existent with a crossover at 4 points between 05:00 to 08:00 and from 14:00 to 14:30, where neutral is higher than negative. Additionally, at various points negative and neutral are mirror opposites which seems unusual as with other graphs above there is more changeability in the sentiment category throughout the event for negative and neutral.

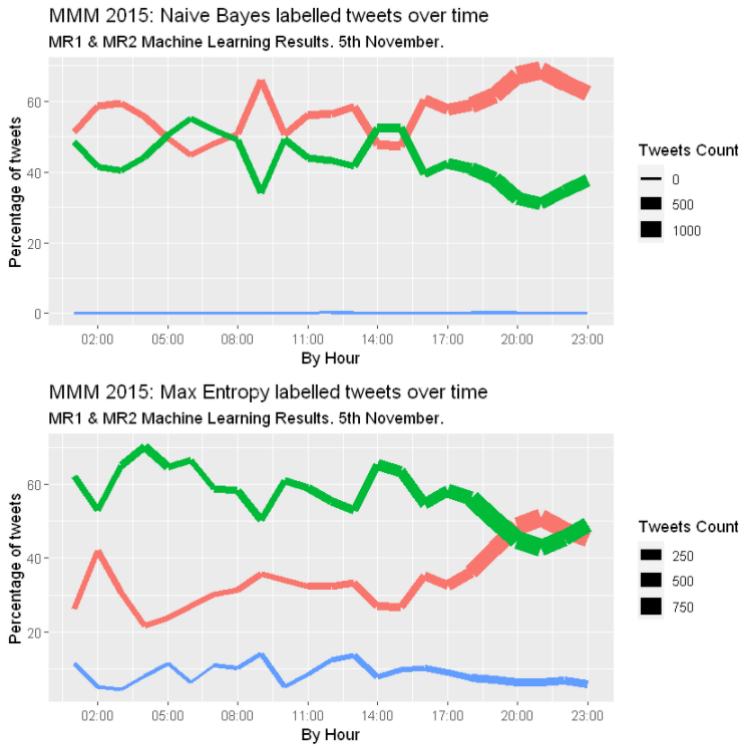

Figure 6.11 2015 MMM - prediction of sentiment by Naive Bayes/ Max Entropy over time (automated)

In Figure 6.11, MaxEnt is different to NB, where neutral and negative crossover at two points at 20:00 and 23:00 and still has mirror oppositive line, but less so compared to NB due to the difference in the count of tweets. The most notable difference to NB is that neutral is majority, positive count is higher (dictionary results in sections 6.5.4 to 6.5.4.2.1 support that Max Entropy identifies more positivity albeit low which shown in Figure 6.11), and the 'Tweets Count' range is narrower than NB. Furthermore, the shape of both green lines in these two plots are similar, whilst the red lines are similar towards the end only and positive in blue is almost not detected by NB and hence yields the low flat line which correlates to the dictionary results (refer to dictionary results in sections 6.5.4 to 6.5.4.2.1) which showed NB was poor in the identification of positive sentiment. Figure 6.11 shares similarity to Figure 6.3 in sentiment trajectory, as there is a mirror opposite with negative and neutral, but the most noticeable difference is negative being the majority. In Figure 6.4, there is again a similar trajectory where it has a crossover and neutral is the highest count which matches MaxEnt. This shows that the predictions are supported by the initial analysis. MaxEnt

has closer resemblance to the previous outlined results, as the result shows a stronger connection than NB.

The thread of results has a reasonably strong connection throughout from the peak time of sentiment classification, average, and changepoint to the predicted results, but MaxEnt showed a higher level of connection than NB when relating to other results. The predicted results where it showed peak or trough aligned with the changepoint and peak sentiment graphs will support the reasons for the trajectory of its negative, neutral and positive categories. In section 6.7.2 we will explore 2016 MMM results and determine any similarities and differences compared to the 2015 MMM.

6.7.2 MMM 2016

In both Figure 6.12 and Figure 6.13, again the results of the majority vote of the sentiment categories are shown by day and hour over a six-day period. There is a total count of 3,356 tweets for manual and 15,491 for automated. These graphs show a similarity with section 5.12.1.1, where the higher peaks of sentiment are shown before and on the day, with neutral being the highest, closely followed by negative and positive.

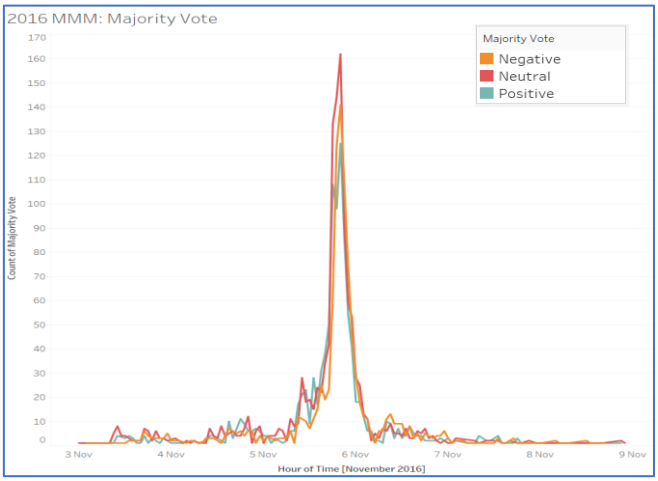

Figure 6.12 2016 MMM sentiment by day/hour (manual)

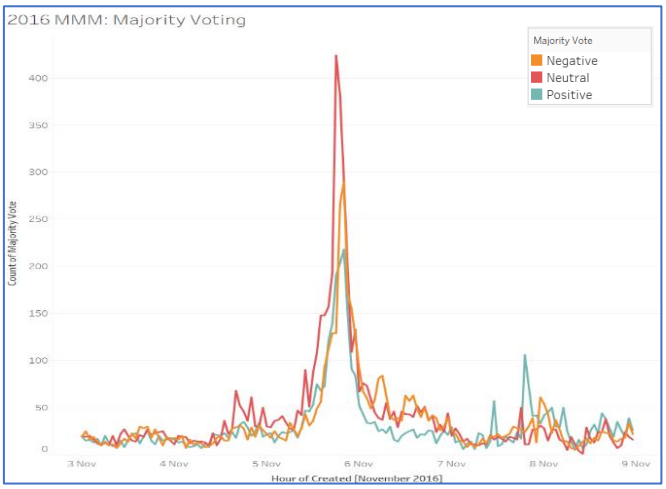

Figure 6.13 2016 MMM sentiment by day/ hour (automated)

During the 2016 MMM certain events are known to have happened and these have been labelled on Figure 6.14 and Figure 6.15, but below provides a detailed description of the events and they can be seen to coincide with some of the peaks and troughs in the sentiment (Gutteridge & Wood, 2016; Nagesh, 2016a; Sims, 2016; The Guardian, 2016): -

- Demonstration began peacefully, with several participants climbing the base of Nelson's column chanting the slogan "one solution, revolution". The procession went along Whitehall, where angry scenes started as it appeared police formed a ring of steel outside parliament. Later, some participants paused to read a message projected in green letters on to a building lining Parliament Square, which said *"Please observe Public Order Act restrictions. Failure to comply may result in arrest and prosecution. Officers may require you to remove facial covering. Failure to comply is an offence."* (Gayle, 2016) This came after members of the crowd ignited fireworks/flares outside Westminster Abbey.
- Metropolitan Police (MET) at 19:00 publicized 10 arrests and a further 33 arrests at 21:00. The demonstration in Parliament Square lessened to several hundred people by 19:30, when one man was seen led away by officers. Splinter groups roamed between Trafalgar Square and Whitehall. One man climbed on top of the memorial to Field Marshal Haig and shouted "this is for all of us" to onlookers.
- Chaotic scenes occurred shortly before 9pm, with riot police moving in to make arrests. A group of protesters surrounded and charged the officers, with shouts of "fuck the police" and "police brutality". Several glass bottles were thrown as police escorted a protester away.
- At 10:45pm, the total amount of arrests is 47, of which the majority are for drug offences and obstruction. MET imposed restrictions to limit the event to a three-hour period between 18:00 and 21:00, and demonstrators are prescribed a route between Trafalgar Square and Whitehall. Additionally, static protests can occur in Trafalgar Square, Richmond Terrace and Parliament Square.

In Figure 6.14, the neutral category displays the highest percentage of tweets throughout the event except when both neutral and positive at 15:00 are 35%. The neutral line peaks twice at 51% at 18:00 and 20:00 which coincides with the bulleted list of events above, where there are marching on streets and police intervention. This is supported by the 'Tweet Count' rising from 18:00 until 21:00 that continues to follow a downwards trajectory. Additionally, the neutral line is more volatile than 2015 MMM, as it has more peaks and troughs with two highest peaks. The negative category is mainly between 20% to 30%, but more positive tweets is higher than negative at the beginning up till 18:30, however, its 'Tweet Count' is lesser than negative in that period.

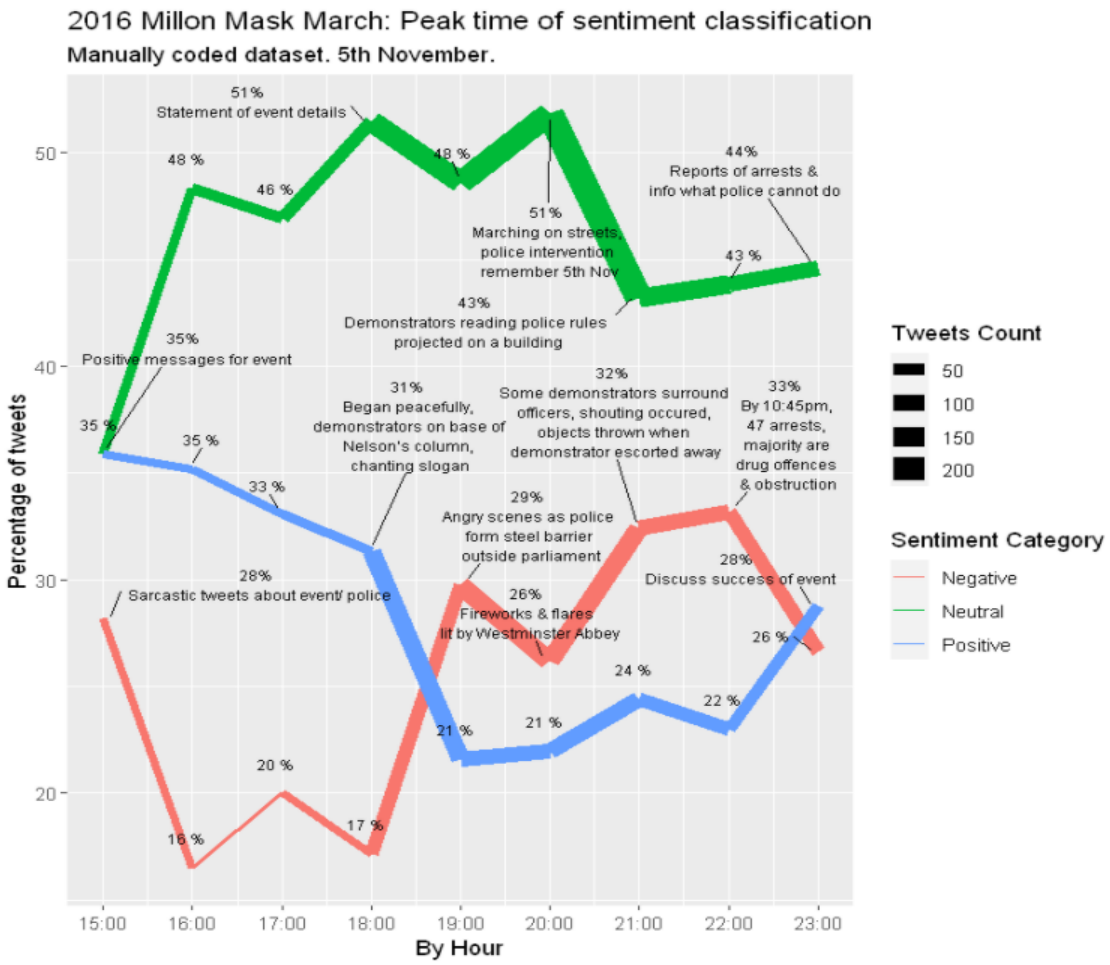

Figure 6.14 MMM 2016: Peak time of sentiment classification (manual)

In Figure 6.14, the negative strand highest peak is above 30%. The first peak is at 22:00 with 33% which matches the bulleted reports related about “number of arrests for specific offences”, and the second is at 21:00 with coincides with reports of demonstrators shouting, objects thrown and conflict with the police. Neutral showed upward turn earlier at 16:00 (which relates to both bulleted reports above and tweets stating event details), but negative rose higher from 18:00 onwards and positive started higher, but then declined at 18:00 until 22:00 where it slightly increased. The positive line at 15:00 has a percentage of 35% positive messages for the event, at 18:00 matches report peacefully chanting slogans, and end of event detailing success of the event on Twitter at 23:00.

In Figure 6.15 it shows the neutral and negative are intertwined from 20:00 to 23:00 with a lower positive share which is expected, but this time interval appears 1 hour less than Figure 6.4. Similarly, before 20:00 was on the opposite scales with neutral’s highest percentage between 30% to 56%, and negative between 17% to 38%, which is mostly higher than 2015 MMM in terms of percentage, however, ‘Tweet Count’ is much less in comparison. The tweets count is far greater in volume than Figure 6.14, which is the reason for the dramatic change in sentiment category trajectories. However, the information outlined at the peaks and troughs is similar to Figure 6.3 despite the sentiment category being apart. Figure 6.15 is comparable to Figure 6.4, as

both neutral and negative are intertwined later than being apart as shown in both Figure 6.3 and Figure 6.14, which again positive being low share is expected.

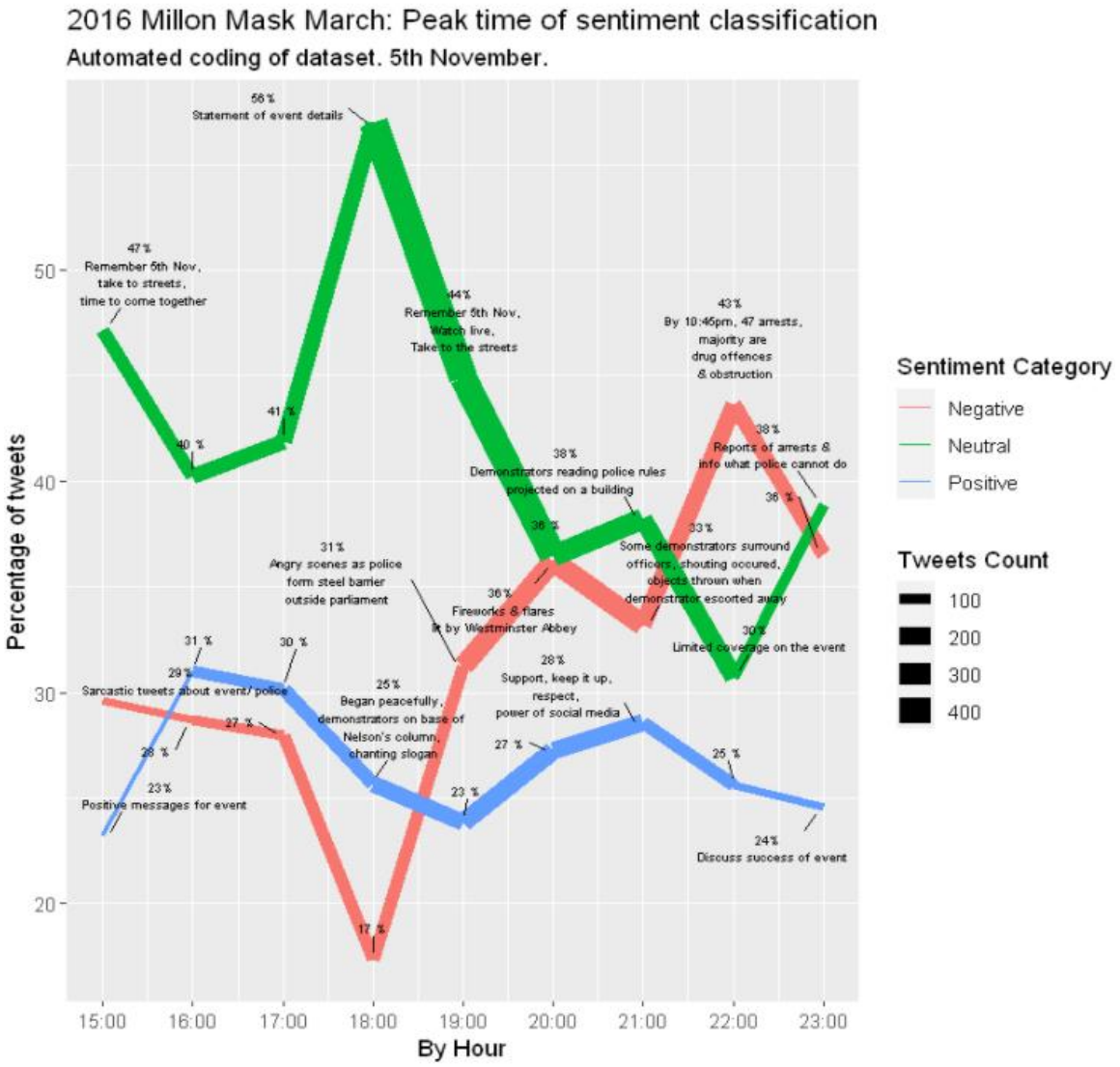

Figure 6.15 MMM 2016: Peak time of sentiment classification (automated)

In Figure 6.15 the neutral line is highest most of the time except at 20:00 when negative and neutral are both on 36%, but at approximately 20:30 negative rises above to 43% and then declines again with a similar percentage before 23:00. At 23:00 neutral is above negative again. The positive line is more consistent from 16:00 onwards to 23:00 despite the steady decline of tweets. These time intervals with the highest peaks and troughs correlate to the bulleted list of events above and mimics Figure 6.14 turn of events.

In Figure 6.16, again shows another way of illustrating the sentiment where the red dots are individual tweet sentiment score and the black line is the average of tweets for a particular time slot on the day of the event. In Figure 6.16, the high volume of tweets is between 18:00 to 21:00, with most tweets between 19:00 to 22:00. There was higher rise in tweets from 17:00 of 115 onwards to 18:00 when the event began. Between 18:00 to 19:00 it was 303, which sees a surge in tweets between 19:00 to

22:00. The peak was between 20:00 and 21:00 with the highest count of tweets on 428. The number of tweets is steady until after 23:00, which decreases by 141 tweets. These time intervals increase or decrease in tweets is expected as correlates with the graphs above. Overall, the average sentiment remained pretty level throughout Figure 6.16, which the average score is nearest to the hour and the scale on the horizontal axis is largely between 0.4 to -0.7.

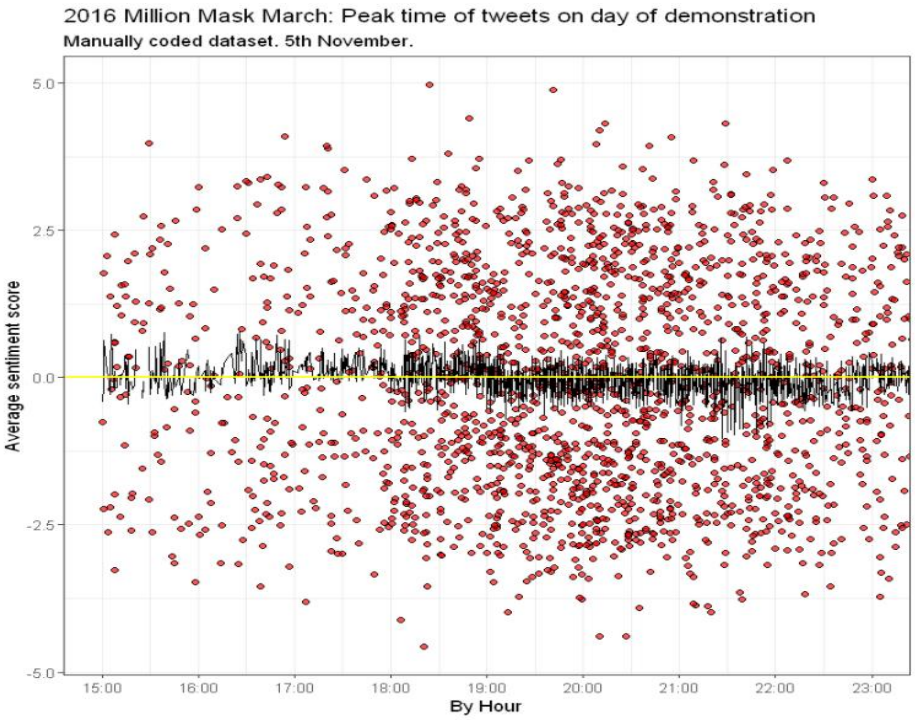

Figure 6.16 2016 MMM - Peak time of tweets on day of demonstration (Manual)

In Figure 6.17, the highest volume of tweets is between 18:00 to 21:00 which agrees with the Figure 6.16 time frame where just proportion is at a higher volume of tweets. The demonstration begun at 18:00, which started to see a higher rise in tweets at 17:00 of 461 tweets, and then from 18:00 to 19:00 saw a significant increase to 744 tweets. The numbers continued rise and its peak was 800 tweets between 20:00 to 21:00, but 21:00 onwards saw a decline in the hundreds with it being small as 143 tweets by midnight. The reason for the surges or decline in tweets is outlined in the timeline events above. Overall, the average sentiment remained near level throughout Figure 6.16, which the average score is nearest to the hour and the scale on the horizontal axis is largely between 0.3 to -0.5 which this range is lower than in Figure 6.16.

[Intentionally Left Blank]

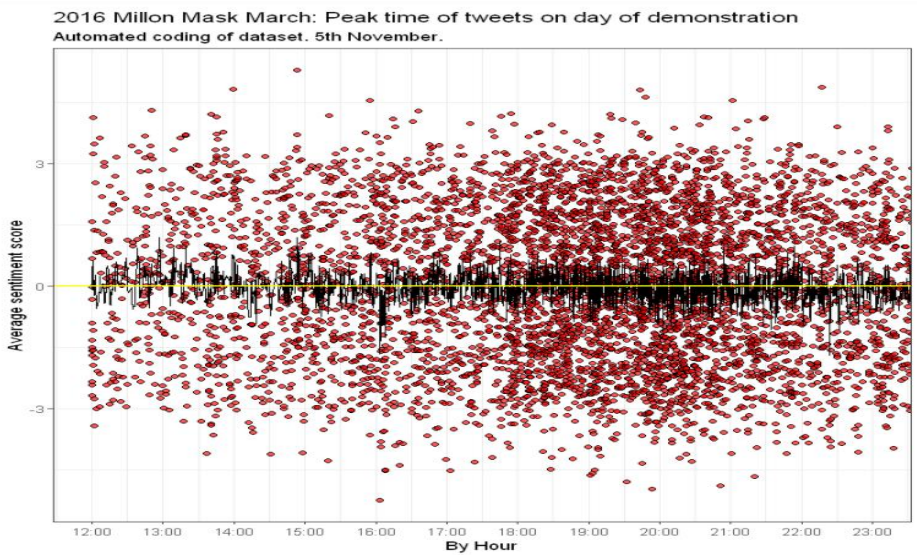

Figure 6.17 2016 MMM - Peak time of tweets on day of demonstration (Automated)

Both Figure 6.16 and Figure 6.17 2016 MMM plots are very similar to the volume of red dots and black average score line for 2015 MMM. In Figure 6.18 provides is a detailed view of the average score for the manually coded (relevant) tweets. As a result, the score is mainly on the negative side except for at 16:00 and 18:45 which shows positive tweets about the event mainly between 0.12 and 0.16. At 19:00 it is -0.05 onwards where the negativity grew over time to a peak of -0.10 at 22:00 which coincides with reports when police containment of chaotic scenes and bottles being thrown. However, at 23:00 it rose to positivity of 0.03 where demonstrators discussed the success of the event.

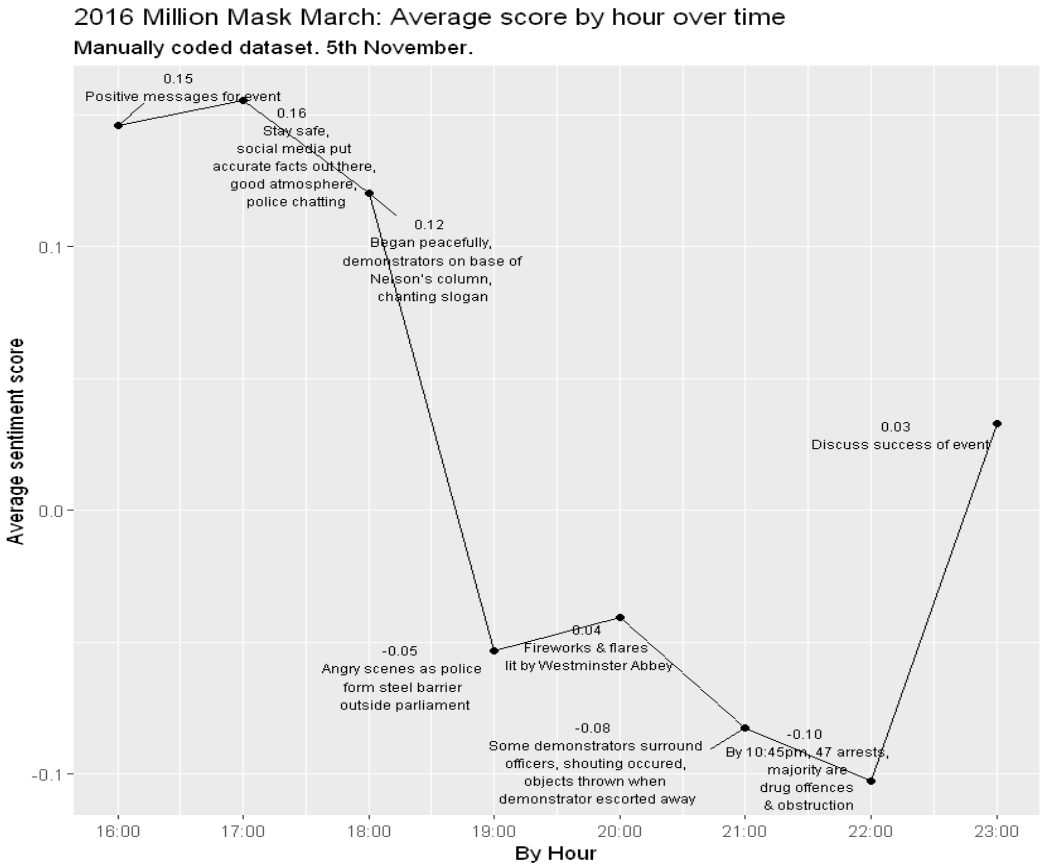

Figure 6.18 2016 MMM - Average score by hour overtime (manual)

Figure 6.19 is similar to Figure 6.18 as line of path over the duration of time, but the average scores are more negative which is mainly due to larger sample of data. The average score leans more on the negative side except for at 16:00 and 18:45 which shows positive tweets about the event mainly between 0.12 and 0.16. At 19:00 it is -0.05 onwards the negativity grew over time to a peak of -0.10 at 22:00 objects are thrown, and demonstrators escorted away by the police which coincides with the events in the graphs above and bulleted list for the description of the timeline. However, at 23:00 it rose to a positivity of 0.03 where demonstrators discussed via Twitter the success of the event. In both Figure 6.7 and Figure 6.8 for 2015 MMM showed high similarity of negative sentiment over time, and Figure 6.19 compared with Figure 6.18 shows the average score follows a similar pattern even though positivity is on a lesser scale, but the main difference is at 17:00 when it starts as negative.

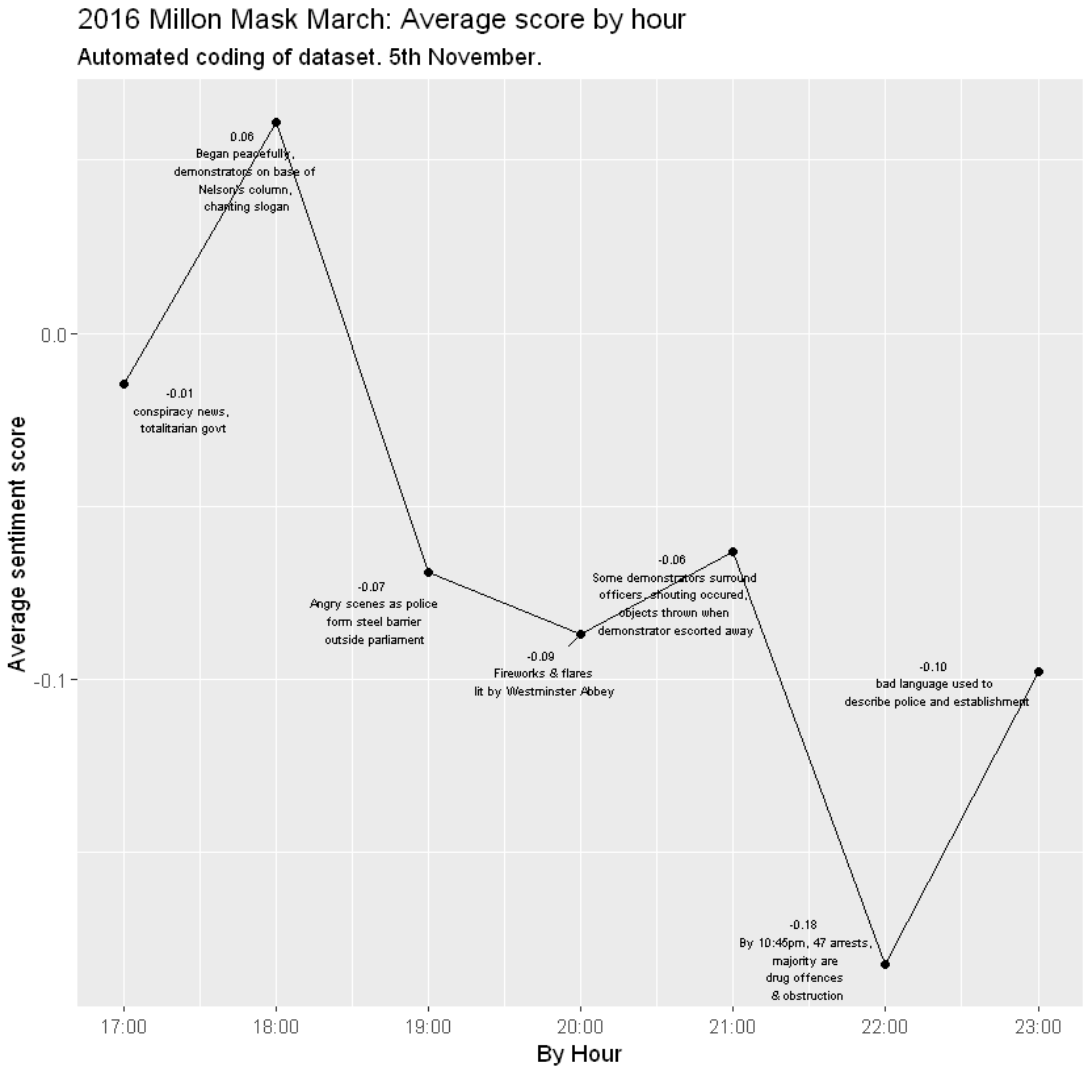

Figure 6.19 2016 MMM - Average score by hour overtime (automated)

Figure 6.19 displays an average score for the automated coded tweets, which shows the score is mainly on the negative side with positivity around 0.06 shown between 17:15 and 18:45. This is similar to Figure 6.18, but shows a higher level of positivity score around 0.15 through nearly the same period from 16:00 to 18:45. At 19:00 it is -

0.07 onwards as the negativity grew over time to a peak at 22:00 of -0.10, but saw less negativity at 21:00 of -0.6, but follows a less similar path to 2015 MMM as negativity decreased towards 23:00 instead of 22:00, which matches when there is a police containment to disperse the demonstrators arrived. Figure 6.19 has a highest negative score of -0.18 (2015 MMM -0.01 less), which is greater than the Figure 6.18 score of -0.10 (2015 MMM -0.13 less).

We have explored the sentiment categories volume over time, average score and density of tweets, which can help to determine the significant occurrences over time. However, in Figure 6.20, again sentiment categories for less than 10 tweets have been excluded, which has impacted the timeline, as there are gaps in the time. The red line is limited to a maximum number of changepoints of 5 as a number must be defined. The negative line for the day of the event has two defined changepoints that are positioned in the morning for 2015 MMM, but 2016 MMM shows four on the day of the event at 16:00, 19:00, 22:00 and 23:00 with one point of the change points on the 6th of November at 01:00 which does not seem to be highly important due to the very small number of tweets and limited information. The major difference is the identified points of change show the highest volume of tweets compared with 2015 MMM and the most significant points of change occur on event day.

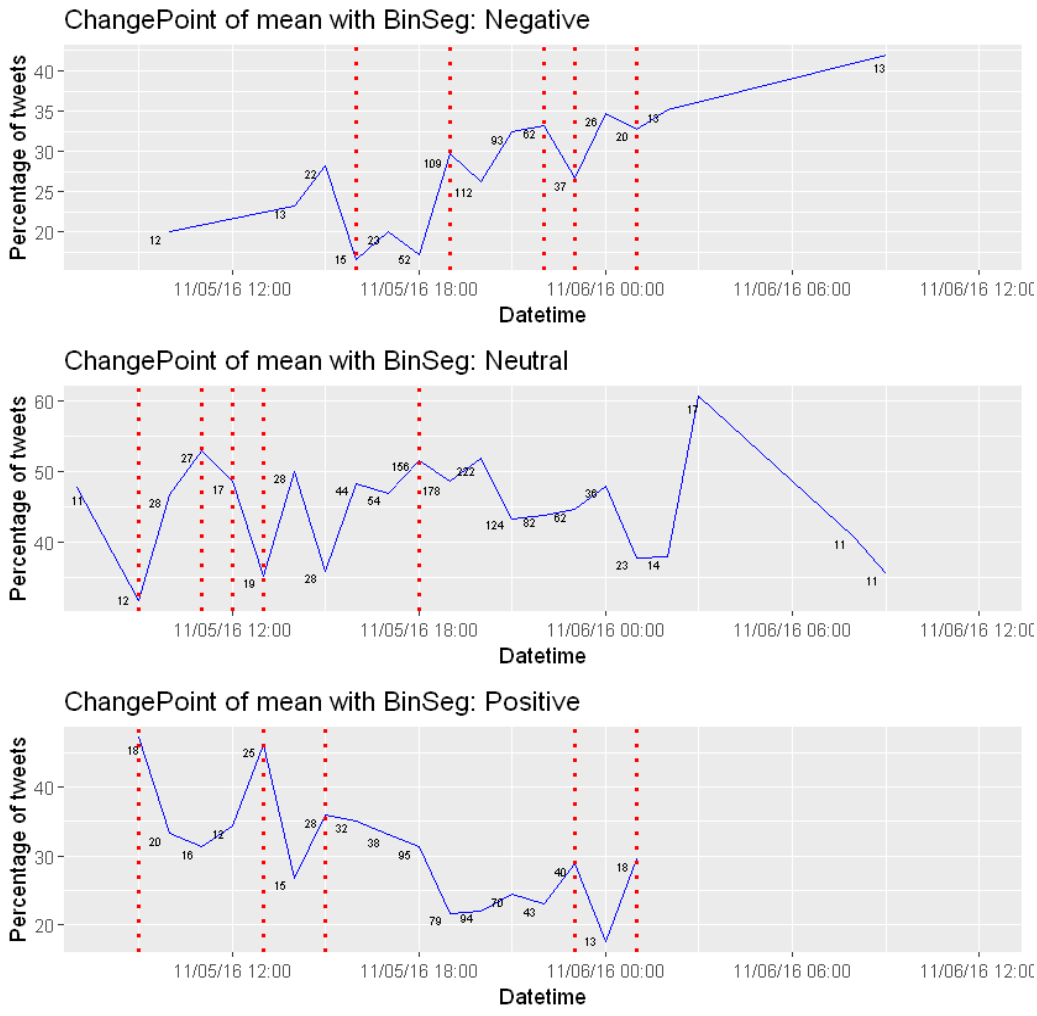

Figure 6.20 ChangePoint of mean with BinSeg by sentiment classification (manual)

In Figure 6.20 there are less tweets identified for each sentiment category before 17:00 on 05/11/16, but for specifically for negativity this grew in volume the time just before the event began at 18:00, and rose even higher during the event. At 16:00 tweets suggest there are negative opinions (many swear words used) aimed towards the police. At 19:00, similar to 2015 MMM, which coincided with the same reports of angry scenes as police formed a steel barrier outside parliament, whereas in 2015 MMM event fireworks were being let off with riot police arriving. At both 22:00 and 23:00 (after the event finished at 21:00) this coincides with reports of further arrests for drug offences, obstruction, and police tried to disperse the remaining crowd and warned conditions of event being violated, and bad language was aimed at the police. On 6th of November at 01:00 (instead of 03:00 like 2015 MMM), it is outlined there was chaos/clashes in central London at the event.

In Figure 6.20 the neutral line change points are that all changepoints are on the event day except 09:00, 11:00, 12:00 and 13:00 are not during the event and contain some of the lowest number of tweets, but at 18:00 shows one of the highest which is expected. It appears for most change points have not been correctly identified. At 09:00, 11:00, 12:00 and 13:00 the tweets are around "remember 5th November, worldwide event and details about where a march is near you, million mask march is tonight, 'guy fawkes' masks, police crackdown this year/security preparations and event goes down in history". At 18:00 there is a "statement about event details, watch live streams of event, take to the streets and at Trafalgar square for the event" which is expected as demonstration would like to gain further traction to highlight their key issues.

In Figure 6.20 the positive line has the lowest volume of tweets. The key change points on the 5th of November are at 08:00, 13:00, 15:00 and 23:00, and on 6th of November at 01:00. At both 08:00 and 13:00 tweets range from being "thankful for something to believe in, thankful for the people behind masks, uniting together for a better future, thank you for support, solidarity for taking part and good luck/stay safe/have fun to those at the event". At 15:00 there are similar positive tweets in support of the event, such as "please be careful/stay safe, happy November 5th march, spread the love and power of truth/trust", and at 23:00 on 06/11/16 at 01:00 the Twitter discussion is of success of the event, for example, "protests were fantastic, good job brothers and sisters in London, thank you to all that turned up and thank you anonymous everywhere". The topic of the positive tweets is similar to 2015 MMM where positive happens to be highlighted before and after event rather during the demonstration.

In Figure 6.21 the automated coded (relevant) sample of tweets displays a more complete time of events than Figure 6.20 due to the larger size dataset. The negative lines for 05/11/16 are at 03:00, 06:00, 08:00 and 20:00, and for 06/11/16 they are 04:00, 05:00, 07:00 and 21:00. On 05/11/16 from 03:00 to 08:00 change points are early in the morning that share some of the lowest tweet counts and there is limited

information provided (such as “demo faces police crackdown, bad language used to describe gov, corruption system is a lie and censoring event”), which seems incorrectly assigned change points. At 20:00 reports coincide with fireworks/flares are let off near Westminster. On 06/11/16 from 04:00 to 07:00 and 21:00 the tweets outline that “activists arrested in London, keep digging the media won’t do it for us and media blackout”. Most of the tweets are more focused on the MMM march in the US with the name of states, presidents, and candidates present in the tweets which is irrelevant to the UK event that project is focused on. Therefore, the automated key words list could be improved to remove tweets not related to the event.

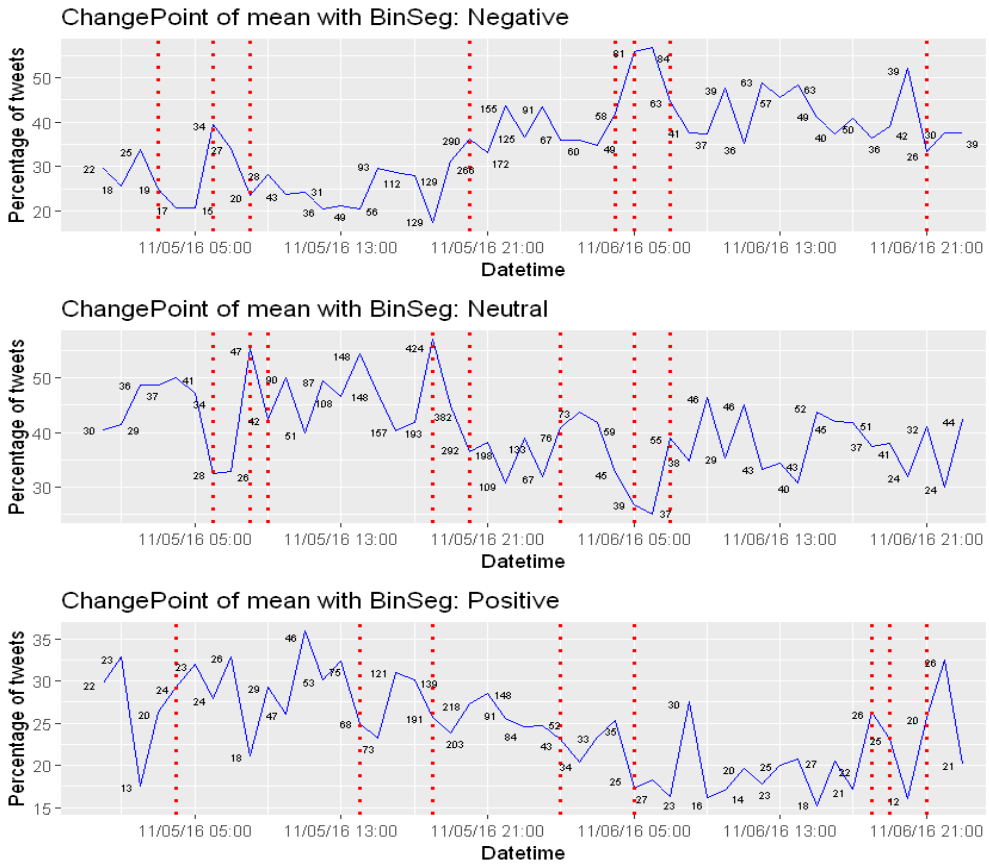

Figure 6.21 ChangePoint of mean with BinSeg by sentiment classification (automated)

In Figure 6.21 the change points for the neutral lines for 05/11/16 are at 06:00, 08:00, 09:00, 18:00, and 20:00 and for 06/11/16 are at 01:00, 05:00 and 07:00. The change points identified at 08:00 to 09:00 (05/11/16) and 01:00 to 07:00 (06/11/16) seem incorrectly identified as change points as low volume of tweets. At 06:00 to 09:00 there are tweets discussion is around “remember the 5th of November, the world is watching, today is the day, million mask march Saturday 5th of November, road closures in Trafalgar Square and thousands expected for annual MMM”. At 18:00 the tweets suggest “take to the streets, live updates from London, remember 5th of November, watch live in London, the march has started to move” and at 20:00 are “take to the streets, rockin’ million mask march in London, who we are and who we are not, remember 5th November and we are moving to Trafalgar Square”. On 6th of November at 01:00 to 07:00 they are “remember 5th of November, we are legion, mmm episode will be out and million mask march across the globe”.

In Figure 6.21 the positive change points for 05/11/16 are 04:00, 14:00, and 18:00 and for 06/11/16 they are 01:00, 05:00, 18:00, 19:00 and 21:00. The change points identified at 04:00 (05/11/16) and 01:00 to 21:00 (06/11/16) seem questionable as change points due to there being a much lower volume of tweets. At 04:00 the tweets suggest "happy November everyone stay safe, happy 5th of November everyone". At 14:00 the tweets see "members are playing spread the love, anonymous call for love, bonfire night lets celebrate, wishing all my buddies in blue a safe night, stay safe" and 18:00 are "the people united will never be defeated, love this truth, good humoured crowd, and hopefully remains peaceful". On 6th of November at both 01:00 and 05:00 it outlined "good job brothers and sisters in London, protest were fantastic see you next year, solidarity with people everywhere, well done to our police officers, thank you anonymous and had a wonderful day". At 18:00, 19:00 and 21:00 they are "more masks please, support other protestors, congrats on successful marches, some members are playing spread the love, and great pleasures in life is doing what others say you cannot do".

Similarly to 2015 MMM, we repeated the analyses for the other techniques used to score the sentiment. As an illustration the machine learning results of the predictions from both Naïve Bayes (NB) and Max Entropy (MaxEnt) are below. In Figure 6.22, the predictions from both NB and MaxEnt display majority for neutral above 40%, with negative mostly between 20% to 40%, and positive on much lower scale up from approximately 5% to 15% of tweets. This correlates to the dictionary results (refer to sections 6.5.4 to 6.5.4.2.1) which showed both NB and MaxEnt (also aligns with the MR1 and MR2 grouped tweets results (refer to sections 6.5.4.1.1 and 6.5.4.2.1) was good in the identification of positive sentiment. There is no crossover between the points as in 2015 MMM, which saw 4 points of a crossover, but both neutral and negative similarly mirror opposite each other with one high and other low at the same time.

[Intentionally Left Blank]

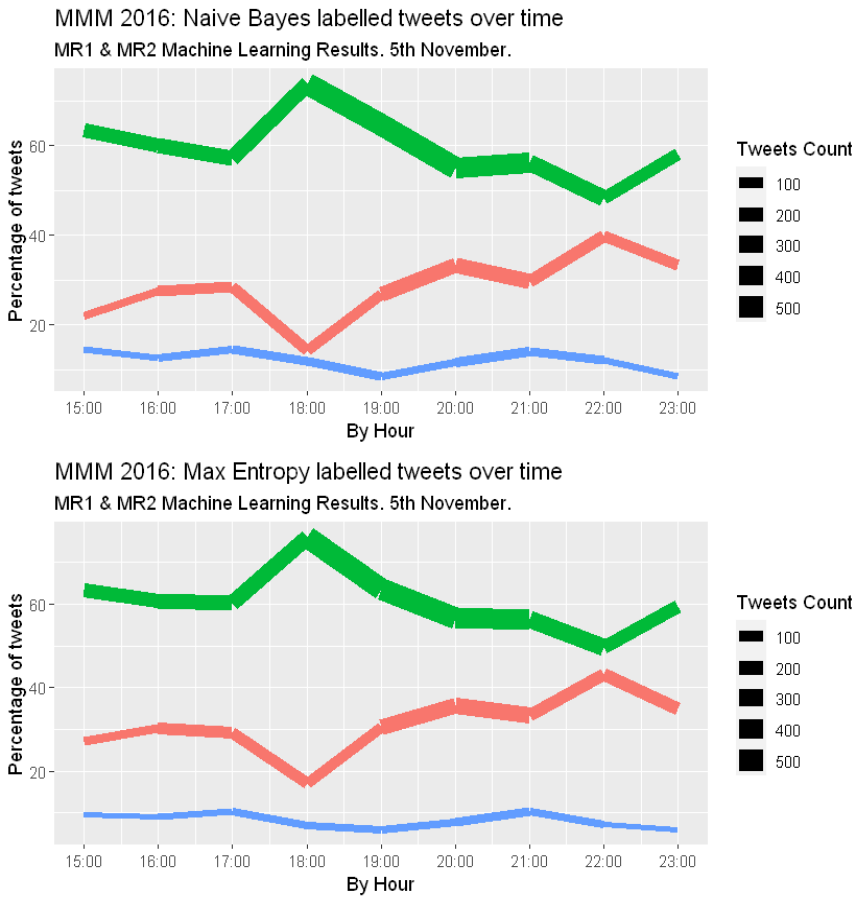

Figure 6.22 2016 MMM - prediction of sentiment by Naive Bayes/Max Entropy over time (automated)

Figure 6.22 shares similarity to Figure 6.14 with neutral with the highest count and negative in behind, with positive higher at the beginning and then has crossovers with negative, with negative rising higher towards 19:00. In Figure 6.17, the sentiment trajectory is more removed from Figure 6.22 than Figure 6.15, as there are multiple crossover points with negative, neutral and positive, but there is some similarity that neutral is highest with negative behind. This shows that the predictions are less supported by the initial analysis, but they have some resemblance due to neutral having the highest count and negative trailing behind. The thread of results has a reasonably strong connection throughout from the peak time of sentiment classification, average, and changepoint to the predicted results, but the predictions are somewhat disconnected relating to the other results. The predicted results where it showed a peak or trough aligned with some of the changepoints for the sentiment categories, but seemed incorrect due to low volume of tweets and limited information which at times was irrelevant, such as tweets relating to US rather than UK for the 2016 MMM event.

In section 6.7.3 we will explore the 2016 Anti-Austerity results and compare them to both MMM events.

6.7.3 Anti-Austerity 2016

The manually coded (relevant) tweets original timestamp for each tweet was incorrect due to the time format not correctly converted in the process of transformation. Therefore, the proportion of times were moved from AM to PM, as the peak was shown at 02:00 to 04:00, which does not make sense as highest peaks are shown in during the event as other event have depicted. The peak time is between 14:00 and 16:00. This time format issue does not effect the automated coded (relevant) data.

In Figure 6.23 show the results of the majority vote of the sentiment categories are shown by day and hour over a six-day period. This contains a total count of 5446 tweets. It is clear there is sparse activity on both 13th and 14th of April, but when it comes to the day before and on the day of the event it rapidly rises, and then declines which is similar to the trend in section 5.12.1.2. In Figure 6.23 the timeline shows neutral has the highest peaks, followed by positive and closely by negative.

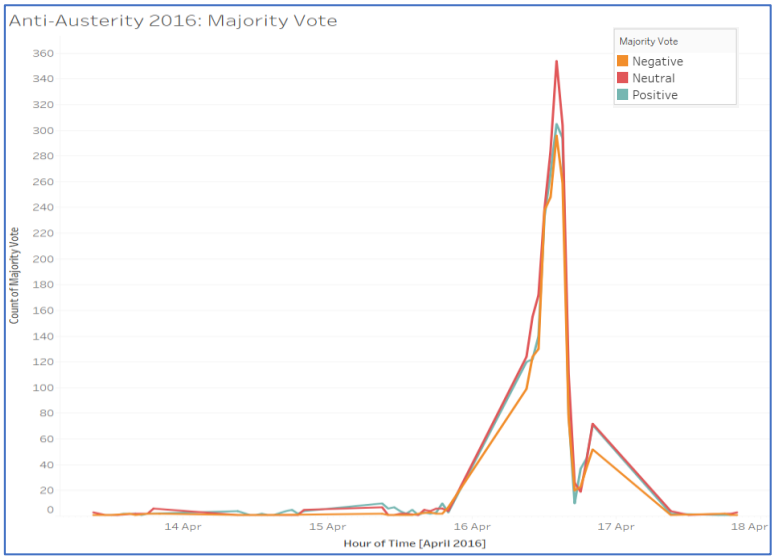

Figure 6.23 2016 AA sentiment by day/hour (manual)

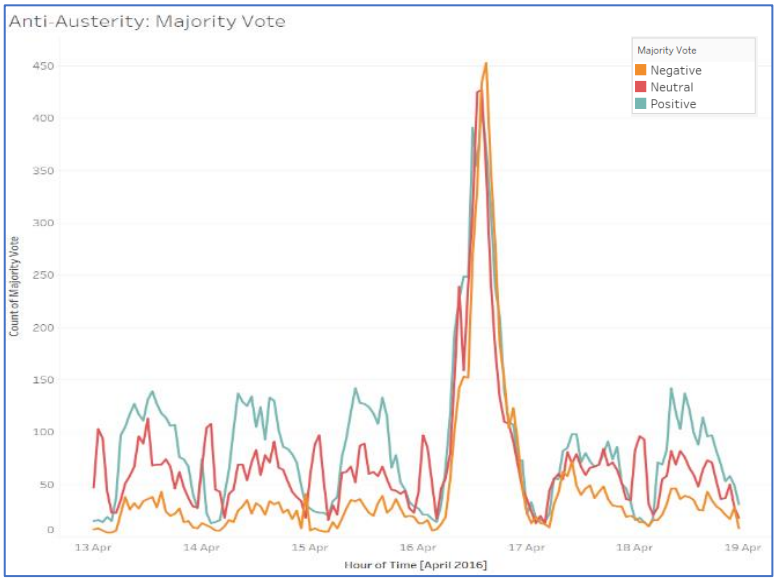

Figure 6.24 2016 AA sentiment by day/hour (automated)

In

Figure 6.24 the total count is 29,963, which mainly shows an incline of tweets and then a rapid decline throughout except when there is a dramatic between 16th to 17th of April, which is not present in Figure 6.23 as has a smaller sample of data.

Figure 6.24 shows results being nearly even at the highest peak for each sentiment category, but both MMM and Dover have a much higher level of both negative and neutral tweets. Additionally, Anti-Austerity shows a higher level of fluctuation, but this may be mainly due to the sample size of a larger dataset compared with the other events. The next stage of the process is to explore the datasets at a greater depth on the day of the event.

During the 2016 Anti-Austerity certain events are known to have happened and these have been labelled on Figure 6.25 and Figure 6.26, but below provides a detailed description of the events and they can be seen to coincide with some of the peaks and troughs in the sentiment: -

- On 16th of April, the National People's Assembly led a national demonstration called "March for Health, Homes, Jobs, Education" (BBC News, 2016a; Grierson, 2016; ITV News, 2016; Unite Community, 2016). Some unions and groups that attended the march included the National Union of Teachers, Stop the War Coalition, the National Union of Students and the Campaign for Nuclear Disarmament (BBC News, 2016a; Grierson, 2016; ITV News, 2016; Unite Community, 2016).
- There was a crowd meeting at Trafalgar Square before the march took place at 1pm. The march began at 13:00 on Gower Street, near the University of Central London, then the demonstrators marched to Trafalgar Square for a rally, where the demonstration lasted till 18:00 (BBC News, 2016a; Grierson, 2016; ITV News, 2016).
- At approximately 15:00, the Shadow Chancellor, John McDonnell, told the crowd a Labour government would end austerity (BBC News, 2016a; Grierson, 2016; ITV News, 2016). Unite's general secretary, Len McCluskey, pulled out a Panama hat during his speech, in a reference to the recent tax scandal, and said: *"The only thing I have from Panama, Mr Cameron, is a hat."* (ITV News, 2016) A video message played to the demonstrators, and the Labour leader Jeremy Corbyn said: *"The austerity we are in is a political choice, not an economic necessity."* (Grierson, 2016) The Green party leader, Natalie Bennett, told the crowd: *"We want all of the Tories out, not just David Cameron. We have a vision of a different kind of society. A society that works for the common good."* (Grierson, 2016)

In Figure 6.25 show the percentage share of neutral, negative and positive tweets throughout the key time period leading up to, and during the event (between 11:00 to 17:00). In Figure 6.25, similar to both MMM events, the neutral category has the highest percentage of tweets throughout the specified duration of the event. However, Anti-Austerity has no overlap with the other sentiment categories as shown

at various times through both MMM events. The neutral line peaks once with 49% at 13:00 when the march began and later it went downwards to 36% at 16:00 when the march ended and then took a sharp rise to 46% at 17:00. Additionally, the neutral line has similar volatility to 2015 MMM with its peaks and troughs. The negative category is mainly between 16% to 32%, but positive is across the board higher in percentage than negative in this period, however, its 'Tweet Count' is smaller than negative, which is similar to 2016 MMM.

In Figure 6.25 the negative strand's highest peak is 32% at 16:00 which coincides with reports for demonstrators voicing their dislike for the Government and awful coverage of the event, and before this it shows a steady incline to 16:00 which further emphasises negativity for the Government and media. The nearer towards 17:00 where the event end there is a sharp drop to 16% which shares dislike for the Government except this time reports tax dodgers and cheats. The negativity in Figure 6.25 is not as consistent with both MMM events, where it shows negativity with a much higher count on most occasions rather than positive as shown for AA.

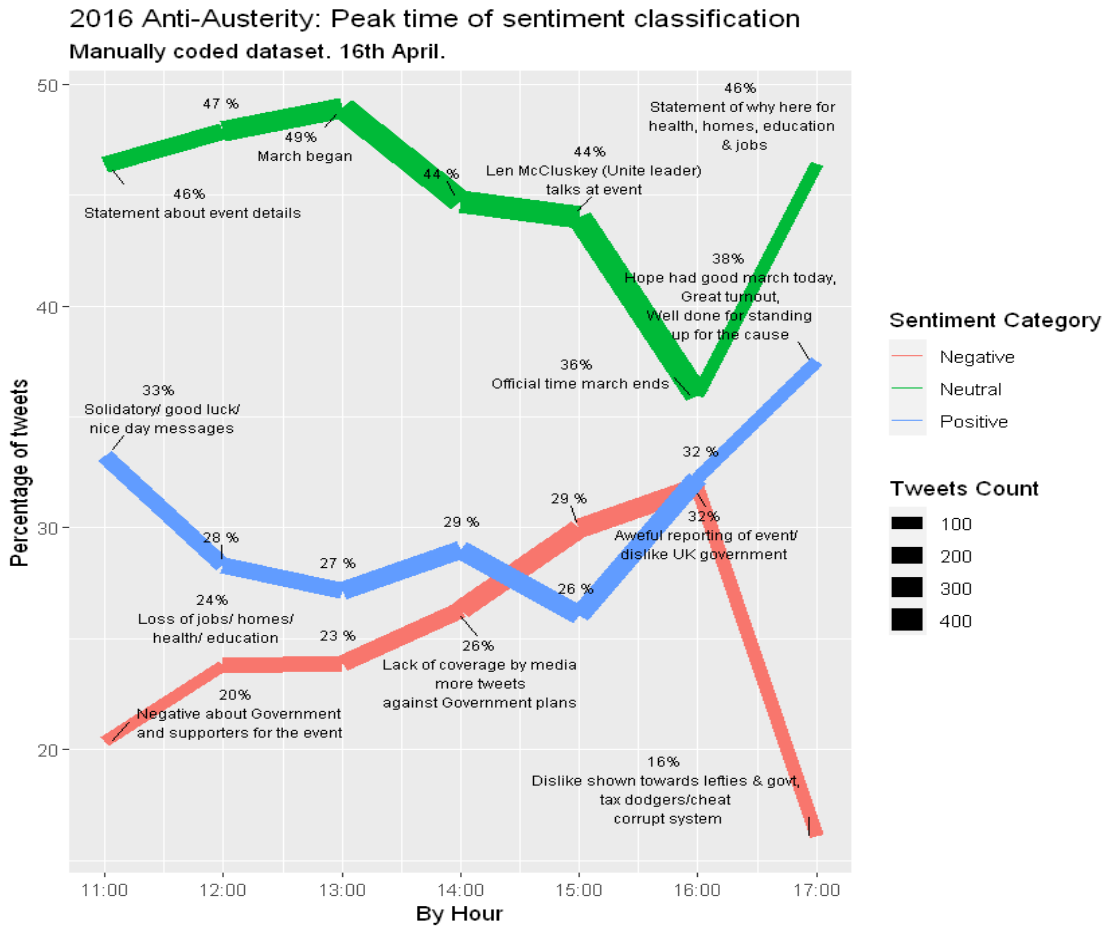

Figure 6.25 2016 Anti-Austerity: Peak time of sentiment classification (manual)

In Figure 6.25 the positive line has its highest peak at 17:00 with 38% positive messages for the event, such as well-done/great turnout. The positivity at the end of the event/ after follows a similar pattern to both MMM events. At 11:00 it's at 33%, which saw a steady decline, then it rose to 29% at 14:00. Towards 15:00 saw a drop by

26% as negativity grew to 29% and positivity increased to 38% by 17:00, which saw a significant drop in negativity.

In Figure 6.26 it shows the neutral, negative and positive are intertwined at multiple points. At 12:30 positive and neutral overlap at approximately 36%, at 13:30 negative and positive overlap at roughly 32%, at 14:00 negative and neutral overlap at estimation of 33%, and at 14:30 positive and neutral are near 32%. In Figure 6.26, there is a higher level of positivity when compared to both MMM events, which leans towards a majority of neutral and negativity. In Figure 6.26, there are sharper differences between their peaks and troughs between negative and neutral compared with positivity which is somewhat more consistent in its trajectory. Additionally, Figure 6.26 is significantly different to Figure 6.25, as it has negative and neutral mirror opposites with positivity higher on the percentage scale, whereas this one shows a greater overlap and higher percentage for each with positivity with the highest percentage level with a slightly lower tweet count. Figure 6.26 shows similar volatility with both Figure 6.6 and Figure 6.17, but those ones had less of a dramatic fall/rise in percentage. Furthermore, manual graphs showed less of an overlap between the sentiment categories than the automated ones.

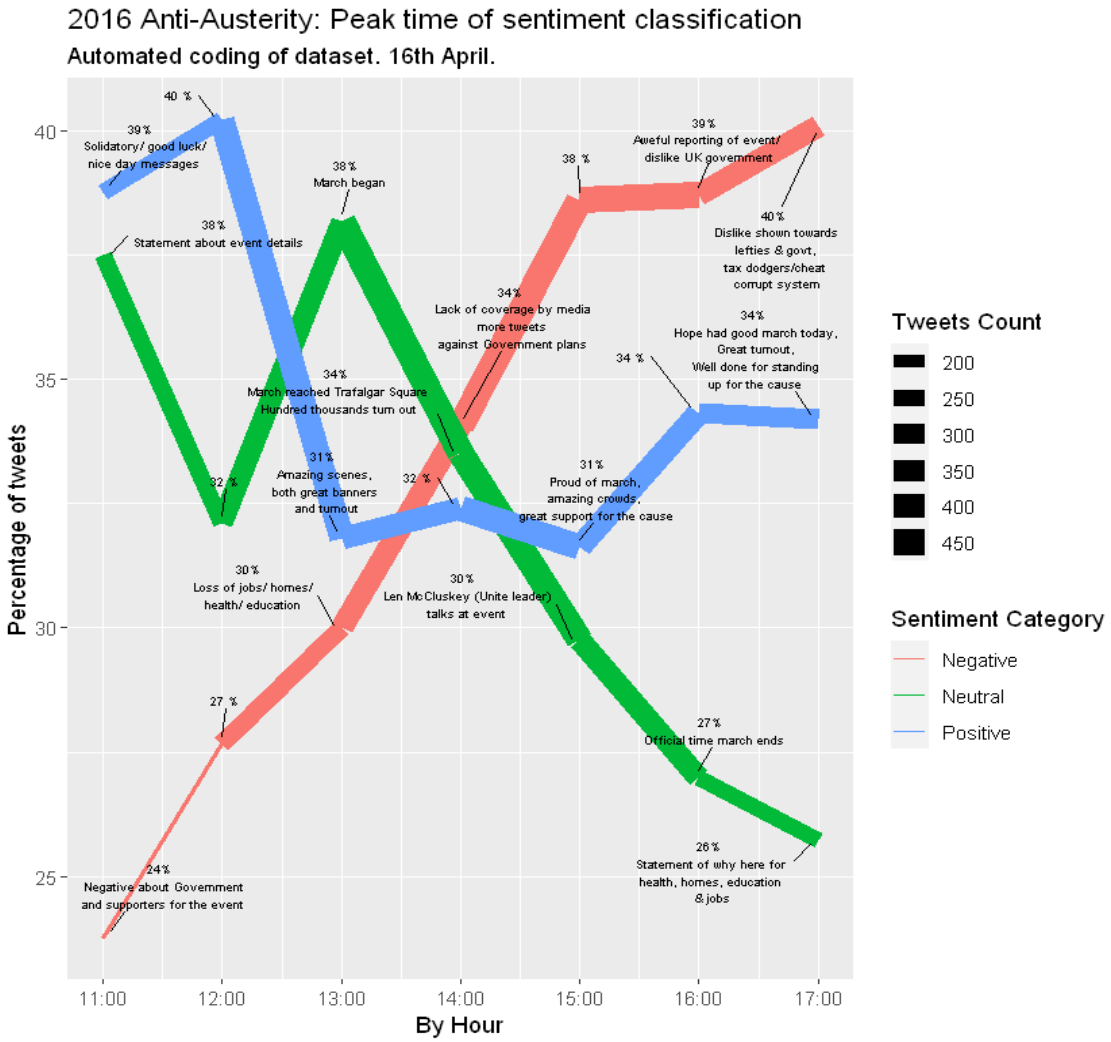

Figure 6.26 2016 Anti-Austerity: Peak time of sentiment classification (automated)

In Figure 6.27 again shows another way of illustrating the sentiment where the red dots are individual tweet sentiment score and the black line is the average of tweets for a particular time slot on the day of the event between 12:00 to 16:00. The highest volume of tweets are displayed between 14:00 to 16:00. There was greater rise of tweets towards 14:00, which kept building onwards to 16:00, and then saw a reduction. The peak volume of tweets was between 14:00 to 15:00. Overall, the average sentiment remained pretty level throughout Figure 6.27, which the average score is nearest to the hour and the scale on the horizontal axis is largely between 0.4 to -0.3.

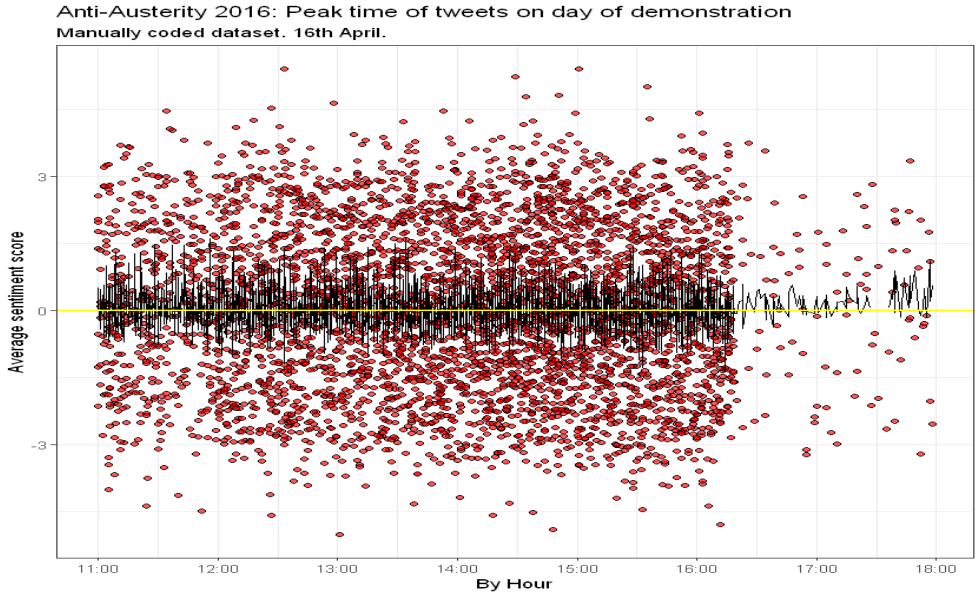

Figure 6.27 2016 Anti-Austerity - Peak time of tweets on day of demonstration (Manual)

In Figure 6.28, the highest volume of tweets is between 13:00 to 16:00 which partially agrees with the Figure 6.27 time frame, just proportion is at a higher volume of tweets. The demonstration begun at 13:00, which started to see a higher rise at 11:00 of 640 tweets, and then 14:00 saw a significant increase to 1,274 tweets. The highest peak is at 14:00, but 15:00 was not far behind on 1,173 tweets, and similar at 16:00 with 912 which coincides with reports of speeches from both Labour and Green party. At 17:00 it is 696 and onwards saw a greater decline. At 18:00 it is 527 tweets. The mains reason for the surges or decline in tweets is outlined in the last two bullet points in the timeline events above. Overall, the average sentiment remained pretty level throughout Figure 6.28 where the average score is nearest to the hour and the scale on the horizontal axis is largely between 0.3 to -0.2.

[Intentionally Left Blank]

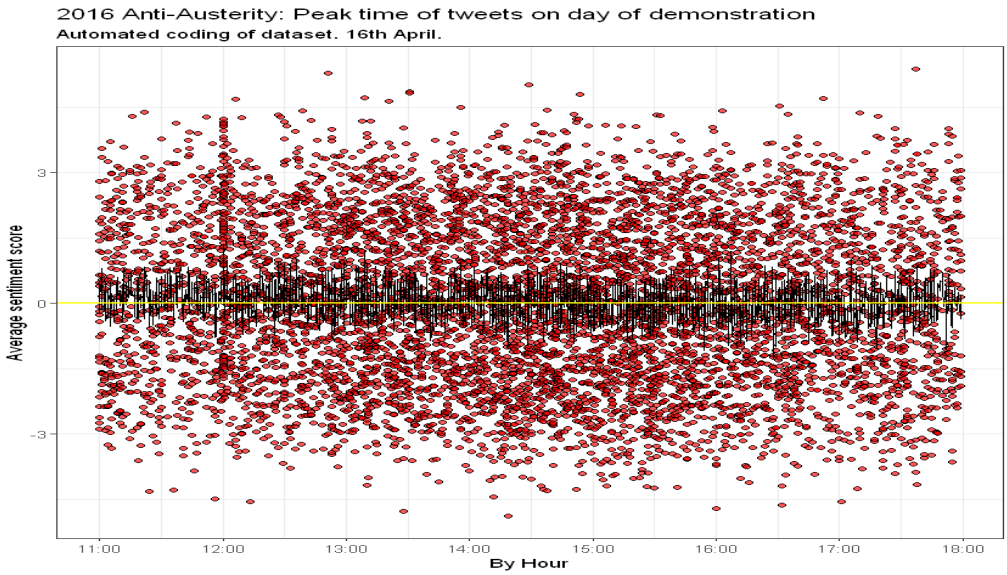

Figure 6.28 2016 Anti-Austerity - Peak time of tweets on day of demonstration (Automated)

Both Figure 6.27 and Figure 6.28 emphasised the volume of tweets and an average across time. However, Figure 6.29 provides a detailed view of the average score for the manual coded tweets between 11:00 and 17:00, which mimics Figure 6.27. Figure 6.29 average score is mainly on the positive side ranging from 0.02 to 0.25 except for at 15:00 on -0.03. At 16:00 it is 0.02 onwards when the positivity grew to a peak of 0.25 at 17:00 where demonstrators discussed the success of the event. This shows the highest level of positivity in an average scored compared to both MMM events.

[Intentionally Left Blank]

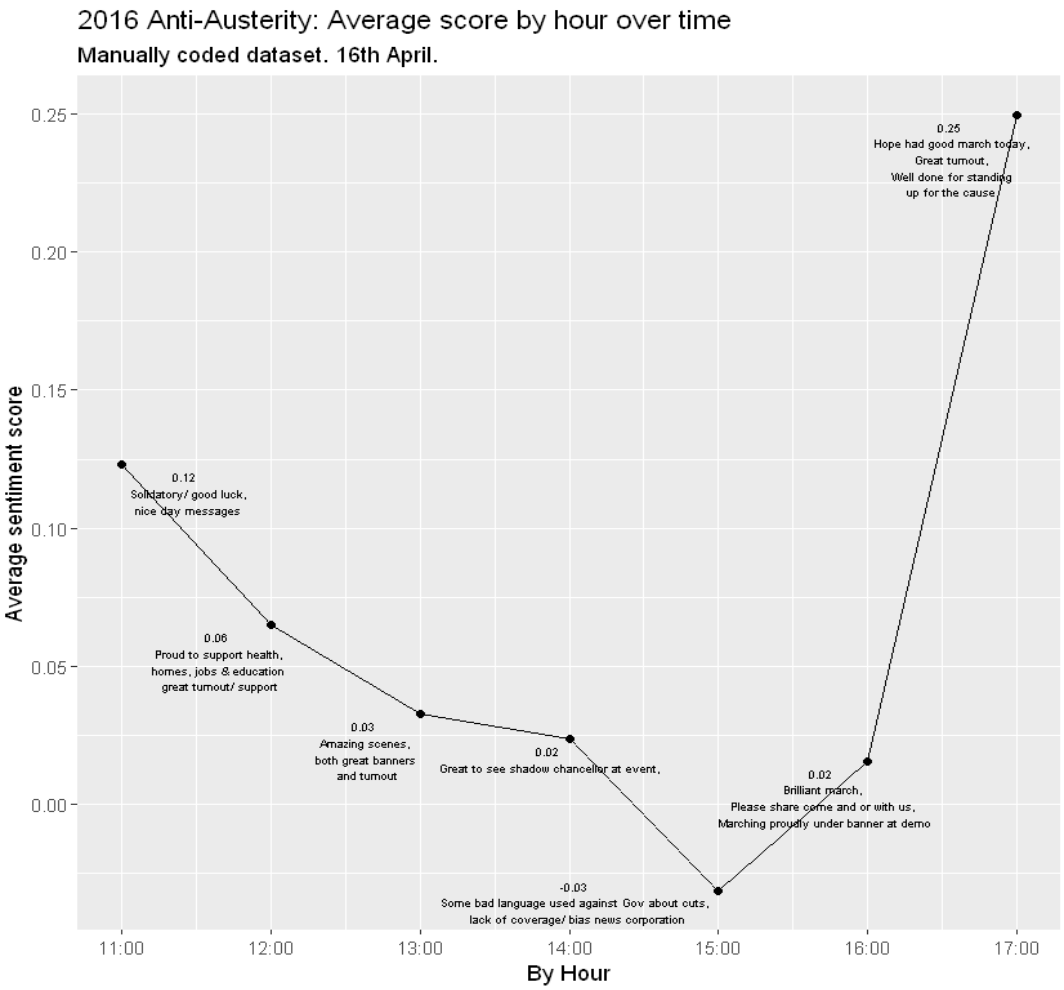

Figure 6.29 2016 Anti-Austerity - Average score by hour overtime (manual)

In Figure 6.30, the average score leans more on the positive side ranging from 0.03 to 0.17 except between 15:00 and 17:00 which is negative between -0.02 and -0.04. The peak for positivity was at 12:00 on 0.17 and for negative at 17:00 on -0.04. This demonstrates that positivity was higher before the event, but when it drew closer to the event negativity grew. At 15:00 two hours into the event it went from average sentiment of positivity to negativity, which grew throughout the rest of the event. Figure 6.30 compared with Figure 6.29 shows the average score follows a similar pattern, even though positivity is on a lesser scale, but the main difference is that it went from positivity to negativity at 15:00 of -0.03, but by 16:00 at 0.02 it was back into a positive average score and the positive peak is at 17:00 rather than before the event as depicted in Figure 6.30.

[Intentionally Left Blank]

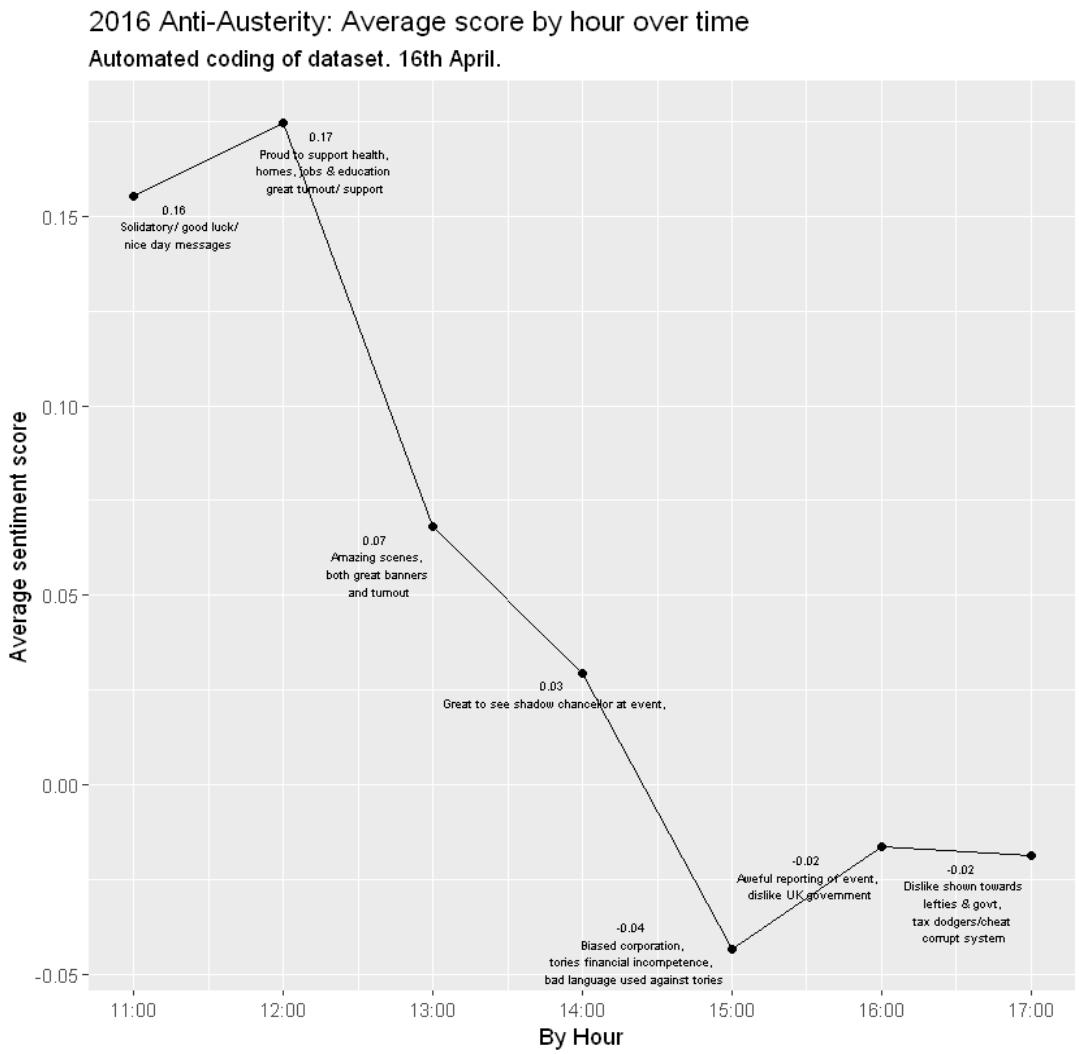

Figure 6.30 2016 Anti-Austerity - Average score by hour overtime (automated)

We have explored the sentiment categories volume over time, average score and density of tweets, which can help to determine the significant occurrences over time. In Figure 6.31, BinSeg is applied in the same way as before to identify changepoints for negative, neutral and positive categories that had more than 10 tweets in the manual coded (relevant) tweets. The red line is limited to a maximum number of changepoints of 5. The negative line has only two change points which are on event day at 15:00 at one of highest peaks and 19:00 at the lowest which appears less significant to due to a much lower level of tweets. This is the first time where the maximum number of change points are not identified compared to both MMM events. At 15:00 coincides with reports when Labour and Green Party speeches took place to the demonstration which sparked a mass of tweets. At 19:00 has the lowest tweets of the day, which tweets suggest is around “bias BBC, PM resign and take party with him, and greed oppression killing UK citizens, and Corbyn cannot hold centre of attention”.

In Figure 6.31 the neutral line has four changepoints on the event day, which are at 12:00, 14:00 (shows the highest volume of tweets on 427), 15:00 and 17:00. At 12:00 “gower street London, so many placards, how to find us for demo, marching with

parents, students and teachers together for education and live stream of demo". At 14:00 change point the tweets suggest 'we are in Trafalgar Square (TS) come and listen to us, assembling/just arriving/filling up at/to TS, at the demo, and rammed at TS" and at 15:00 '150k marched today, John McDonnell at TS now, live feed from demonstrator, here we are in TS'. The tweets suggestion at 15:00 coincides with news reports as emphasised in the second to last bullet point in the timeline of events. When the tweets were closely examined between 14:00 and 15:00, we noticed that many tweets are misclassified negative and should be either positive or neutral. Therefore, more keywords are required to add to find the most relevant tweets and discard more unrelates ones to improve accuracy of the results. At 17:00 the tweets suggest we "can watch demo live, nothing on BBC or SKY tv new channels, TS filled out in all directions, many thousands marching and masses take to streets".

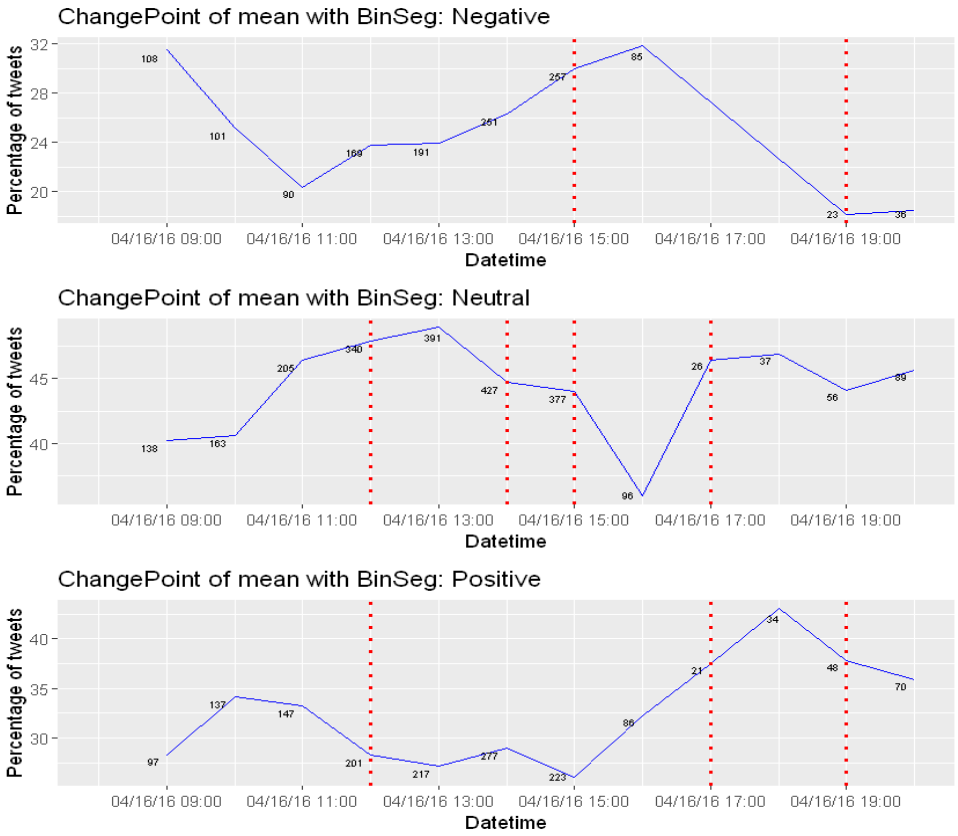

Figure 6.31 ChangePoint of mean with BinSeg by sentiment classification (manual)

In Figure 6.31 the positive line has the least volume of tweets and the changepoints on the 16th of April are only at 12:00, 17:00 and 19:00. Both 17:00 and 19:00 are couple of the lowest number of tweets, but this may have been identified by BinSeg as the positive line is higher for fewer tweets compared to time before it, as percentage of proportion for positive is weighed against both negative and neutral. At 12:00 the tweets highlight being "proud to support the demonstration and great turnout", and at 17:00 "great turnout, well done for standing up for the cause, and hope had good march today". At 19:00 tweets suggest to "join tide of humanity in London, thank you everyone, good on everyone joined the protest today and so much support Corbyn".

In Figure 6.32, the automated coded (relevant) tweets suggest the negative line change points are 16/04/16 at 07:00, 13:00, 15:00 and 23:00 and on 17/04/16 they are 02:00, 03:00, 11:00 and 18:00. The change points identified at early morning times for 16th and all the times for 17th of April tend be very low number of tweets which again may have been incorrectly identified by BinSeg due to how percentage of proportion for negative is weighed against both positive and neutral. At 07:00 Twitter users suggest "MPs refusing to answer questions in parliament, whole system rotten to the core and shout for no confidence" and at 13:00 "corporate tax dodging costs us, anoraks against austerity, tory government is revolting, and angry about media blackout". At 15:00 the tweets suggest we see "bad language used to describe Tories, about Tories financial incompetence, and biased media" which this surge of tweets coincides with reports where both Labour and Green parties are showing support for demonstrators to go against the Conservative party. At 23:00 the tweets further suggest "propaganda corporation pushing gov agenda, biased BBC and all need to stop paying for license fee, and never trust a story". On 17th of April at both 02:00 and 03:00 outlined a "despise for the BBC, bad taste headline and falsified crowd figures and Cameron should resign". At 11:00 tweets suggest the "BBC omissions of late confirm their conflict of interest, media failing to cover event, system is the problem, and demanding PM resign". At 18:00 the tweets show "shame on the beeb, political bias and moaning about austerity should be challenged for mass immigration impact".

In Figure 6.32 the neutral line change points for 16/04/16 are 04:00, 06:00 and 15:00 and on 17/04/16 they are 00:00, 01:00, 05:00, 06:00 and 08:00. Similar to manual coded changepoint the time intervals for early morning contain the lowest tweets which makes it appear less significant. This is further supported on 16th of April at 04:00 the tweets in the sample are irrelevant for the event. And at 06:00 as there are limited tweets that suggest 'early start, get a coach to the march, on route to demonstration, we will see you there at event, and going to be in the march'. However, there are a larger number of tweets at 15:00 that show "listening to junior doctors, scenes from TS, 150K people marched today and with john mcdonnell at TS now" which coincides with news reports as outlined in the timeline of events at the beginning of this section. On 17/04/16 at 00:00 again a small number of tweets suggest 'tens of thousands rally, we want a publicly owned NHS" and at 01:00 again none of tweets are relevant for the event. At 05:00 limited number of tweets outline "see you there and waiting for you here", and at 06:00 a small number of tweets suggest "anti-march in central London, where was the march, why has BBC downplayed the story and future to believe in look around' and lastly at 08:00 'sang along yesterday, part of the media cover up and people power'.

[Intentionally Left Blank]

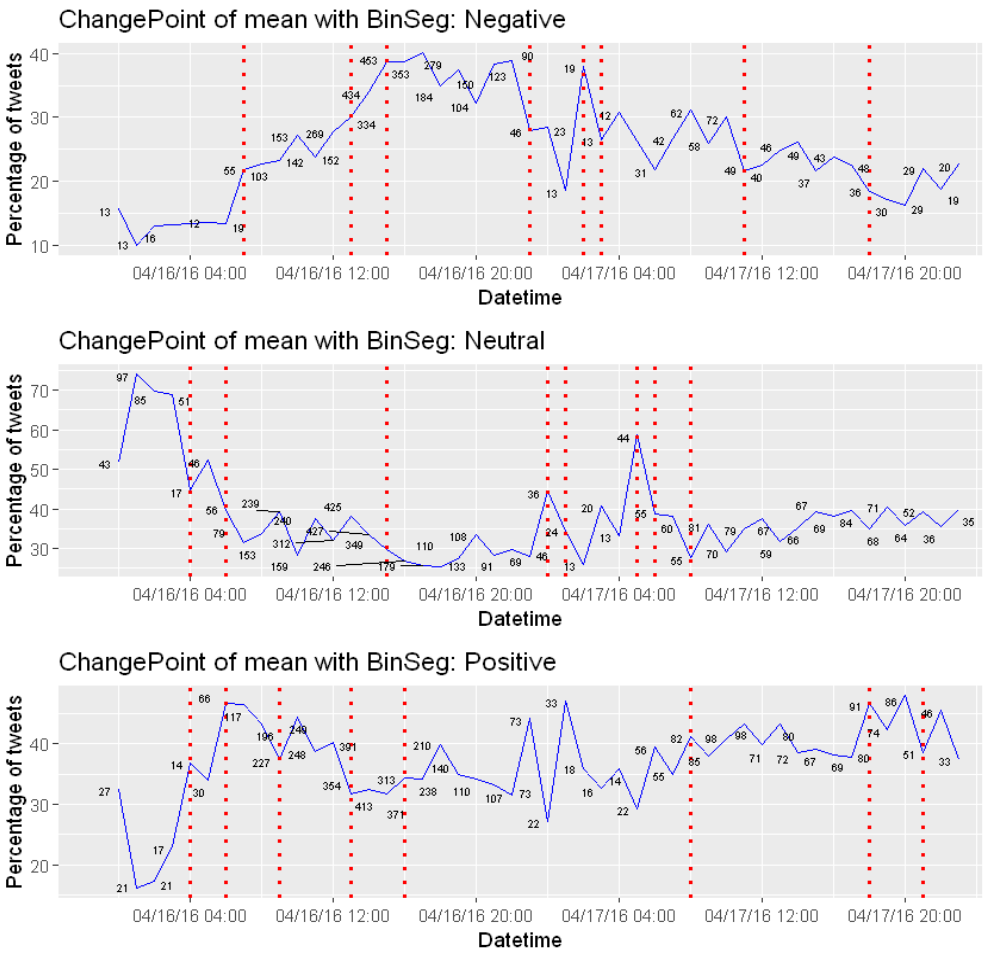

Figure 6.32 ChangePoint of mean with BinSeg by sentiment classification (automated)

In Figure 6.32 the positive line changepoints are 16/04/2016 at 04:00, 06:00, 09:00, 13:00 and 16:00 and on 17/04/2016 at 08:00, 18:00 and 21:00. Similar to manual coded changepoint the time intervals and Figure 6.32 for early morning times and after the event contain the lowest volume of tweets which makes it appear less significant. This is further supported at 04:00 as none of the tweets are relevant, however, at 06:00 is larger number of tweets that are about “wish I could be there, march provides hope to come together to unite against the gov, police policing peaceful demo and setting off to join people for march, and good luck/big support to everyone for the march today”. At 09:00 tweets further suggest a “march together/strike together to win, look forward to joining everyone, thank you to all students/young people making effort and be safe/god love you” and at 13:00 say “thank you to everyone who was there supporting, proud of you all, trending in UK but no mention of event, thank you labour for support, incredible turnout and brilliant atmosphere”. At 16:00 tweets outline “marching proudly, brilliant march, amazing turnout and fantastic/fabulous labour speech”. On 17/04/2016 at 08:00 the limited tweets are about “colourful piece on yesterday demo, which includes a socialist worker popular tweets, great speeches, in solidarity and rapidly growing movement”, and 18:00 a small number of tweets outline “great day meeting new faces and proud of students performance, proud to wave unite flag and great to march with student nurses’ and lastly at 21:00 limited tweets suggest ‘brilliant work during demo and thank you for the support”.

In Figure 6.33, the predictions from both Naïve Bayes (NB) and Max Entropy (MaxEnt) display majority for neutral above 60%, with negative mostly between 5% to 20%, and positive on a higher scale approximately between 20% to 30% of tweets. There is no crossover between the points, which is the same as 2016 MMM and similarly neutral and negative mirror opposite each other with one high and other low at the same time.

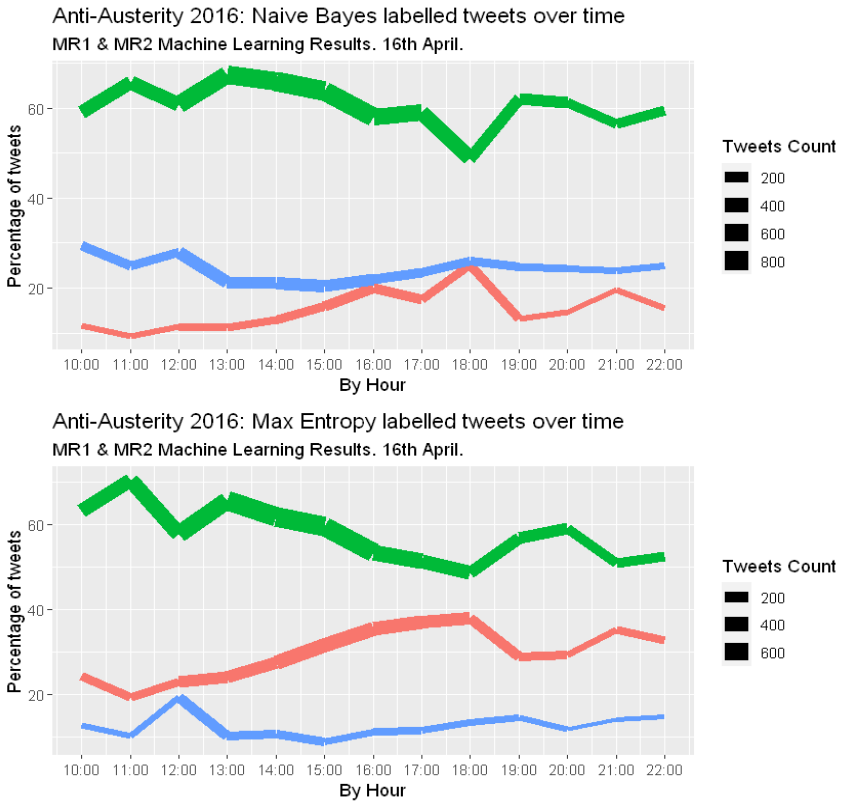

Figure 6.33 2016 Anti-Austerity - prediction of sentiment by NB/MaxEnt over time (automated)

In Figure 6.33, Naïve Bayes result shares similarity to Figure 6.25 with neutral with the highest count and positive in behind, with negative on a similar percentage at 18:00. However, MaxEnt on this occasion has a lower positivity line compared to Naïve Bayes higher percentage of proportion with a higher tweet count, which MaxEnt has a less strong comparison for both Figure 6.25 and Figure 6.26. These algorithms results to identify that positivity correlates to the dictionary results (refer to sections 6.5.4.1.1 and 6.5.4.3.1) which showed MaxEnt (similar for the machine learning results for tweet and manual in sections 6.5.4.1.1 and 6.6.1.1) was less good compared to Naïve Bayes in the identification of positive sentiment. The thread of results has a reasonably strong connection throughout from the peak time of sentiment classification, average, and changepoint to the predicted results. The predicted results are where it showed a peak or trough aligned with some of the changepoints and sentiment graphs that align well with the bulleted timeline of events above and the tweets that indicate the change of trajectory of its different sentiment categories.

In section 6.74 the results from both MMM and AA have been compared against the 2016 Dover results.

6.7.4 Dover 2016

The manually coded (relevant) tweets original timestamp for each tweet was incorrect due to the time format not correctly converted in the process of transformation. Therefore, the proportion of times were moved from AM to PM, as the peak was shown at 02:00 to 08:00, which does not make sense as highest peaks are shown in during the event as other event have depicted. The peak volume of tweets is between 14:00 to 20:00. This time format issue does not effect the automated coded (relevant) data.

Both Figure 6.34 and

Figure 6.35, the results of the majority vote of the sentiment categories are shown by day and hour over a seven-day period. Figure 6.34 contains a total count of 2830 tweets for manual and

Figure 6.35 it is 3174 for automated. These results show a similar pattern to section 5.12.1.3 in their higher peaks of sentiment shown before and on day, with neutral being the highest, closely followed by negative and positive.

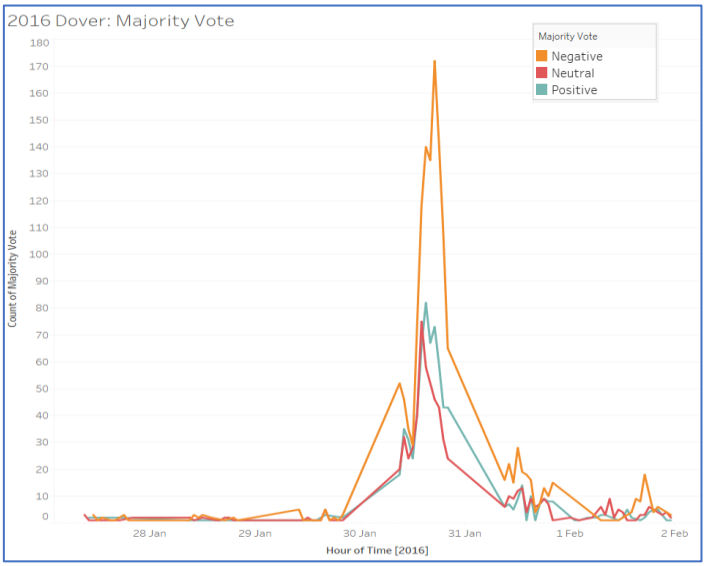

Figure 6.34 2016 Dover sentiment by day/hour (manual)

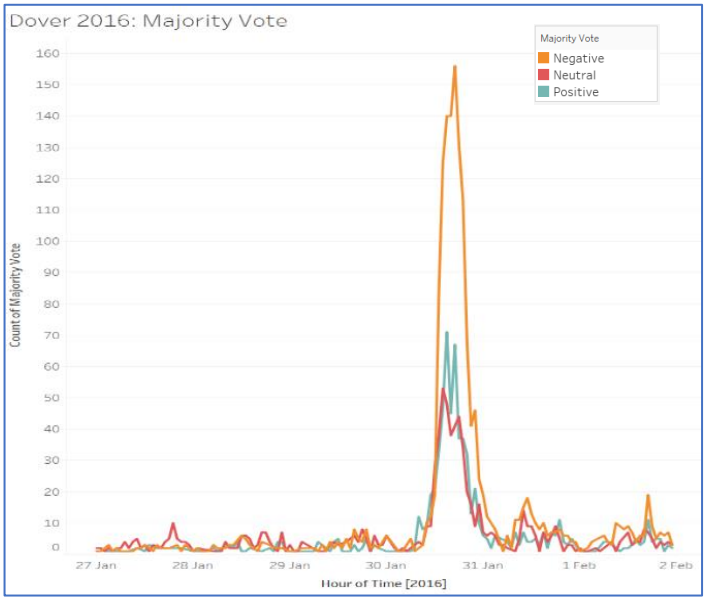

Figure 6.35 2016 Dover sentiment by day/hour (automated)

During the 2016 Dover certain events are known to have happened and these have been labelled on Figure 6.36 and Figure 6.37, but below provides a detailed description of the events and they can be seen to coincide with some of the peaks and troughs in the sentiment (Gayle, 2016; Lennon, 2017; Osborne, 2016; Rkaina, 2016): -

- Around 11:00 violence linked to the demonstrations began at service station before arrival to Dover, when far-right and anti-fascist protesters inadvertently stopped at the same services. The anti-fascist coach windscreen was smashed, and a swastika was drawn on it. Police arrested six people on suspicion of violent disorder.
- Different left-wing groups were held at the Market Square in Dover's town centre at 11:00. A group of masked anti-fascists broke off from the Market Square at 12:30 towards the train station where far-right groups gathered. Far-right demonstration began at Dover's Priory railway station at 13:00.
- Police arrested many demonstrators when the far-right 'East Kent Alliance' managed to clash with their rival group 'Kent Anti-Racism' network violently. At approximately after 13:30, witnesses outline far-right protesters attacked using 'metal poles, sticks and bottles' and anti-fascists hit back with 'bricks and flares' until police separated both groups close to Dover Priory station.
- Police separated the groups and the demonstrators marched through Dover to a rallying point close to the docks where they listened to speeches. Diane Abbot speech for opposing the anti-immigration begins around 13:50 and far-right speeches against immigration takes place near/ around 14:31 which finish around 14:54. After the speeches around 15:35 a far-right demonstrator squares up to a police officer.
- At around 16:00, anti-fascists march back to Market Square in the centre of Dover and the far-right appears to have finished after being escorted back to the station. Reports of nine people were arrested at the event for possession of offensive weapons, breaching the peace, violent disorder and a range of public order offences.

In Figure 6.36, the results are similar to both MMM and Anti-Austerity events, where neutral has a high percentage, but instead negative is more dominate for the duration of the event. Furthermore, Dover is similar to both MMM events as there is an overlap between the sentiment categories. In Figure 6.36 the neutral line peaks at 12:00 on 53% and when the march began at 13:00 it dropped by 14% to 35%, where it stayed in the region of 31% to 37%. The neutral category is the most consistent of the three categories in terms of percentage (except for at 12:00) but compared to the other 3 events the overall volatility has been far greater for Dover. The neutral line overlaps with the negative category at 13:00, 13:30 and approximately 14:15 which both categories at those times have approximately 40% each of the polarity count with a total of 80% combined with 20% for positive.

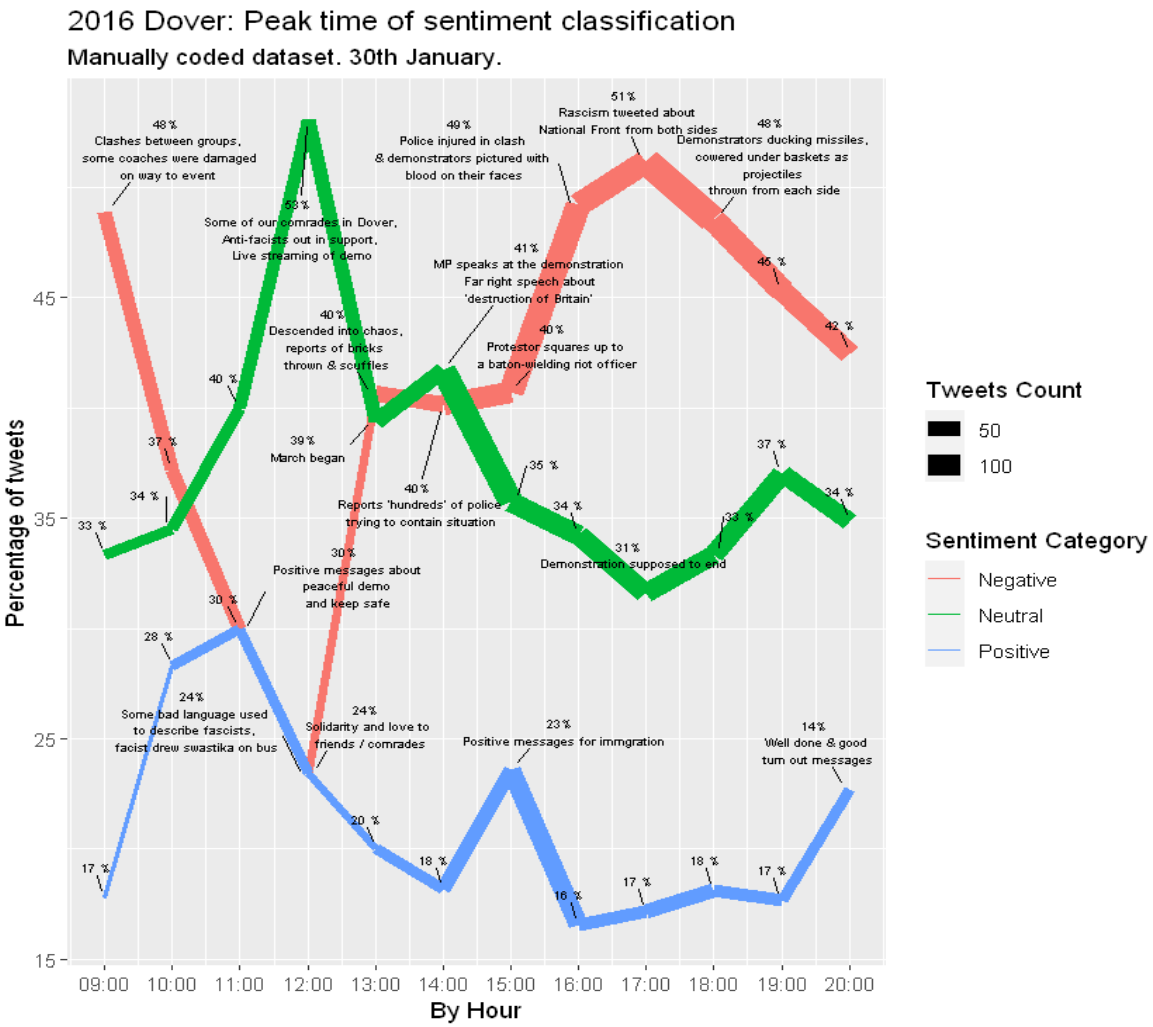

Figure 6.36 2016 Dover: Peak time of sentiment classification (manual)

In Figure 6.36 the negative strand highest peak is 51% at 17:00 which highlights that racism is the topic of debate, and after this time there is a decline of tweets to 42% by 20:00, which further emphasises disruption with missiles being thrown. At 12:00 on 24% is the lowest peak with tweets suggest the use of bad language to describe fascists and a swastika drawn on a bus, which was reported by news media around 11:00. At 12:00 this time overlaps with positivity tweets outline there is 'solidarity is shown to friends/comrades'. The negative line has the largest number of tweets which coincides with the news reports that outlined a wide range of negative issues impacted the event.

In Figure 6.36 the positive lines contain the least tweets counted throughout the event. The highest peak is at 11:00 on 30% positive tweets for the event, such as 'love to friends/comrades'. At 09:00 it is on 17%, which saw a steady incline where it rose to 28% at 10:00, and then saw a drop by 13:00 on 20% as negativity grew to 40%. The range of percentage is between 16% to 23% with most under 20% of tweets. The highest volume of tweets is between 14:00 and 16:00 for positivity, between 14:00 and 15:00 is where the speeches begun for Labour and the far-right group, which could be the cause for the increase as reported in the timeline above.

In Figure 6.37 the automated (relevant) data shows that neutral, negative and positive are on a higher level of points intertwined throughout the event when compared to both MMM and Anti-Austerity demonstrations. There are 5 points of overlap between neutral and positive, and then two for negative where it intersects with the other two sentiment categories. Dover has shown a greater level of negativity than neutral compared to the other events. This may be due to the higher level of disruption in this event which is emphasised in both news reports and tweets compared to the other 3 demonstrations. Additionally, Figure 6.37 is significantly different to Figure 6.36, as there is a greater level of negativity with positive and neutral closely aligned in its percentages of the tweets and there is a near mirror oppositive between the negative and positive line from 15:00 to 20:00 on the percentage scale, but with a greater volume of negativity.

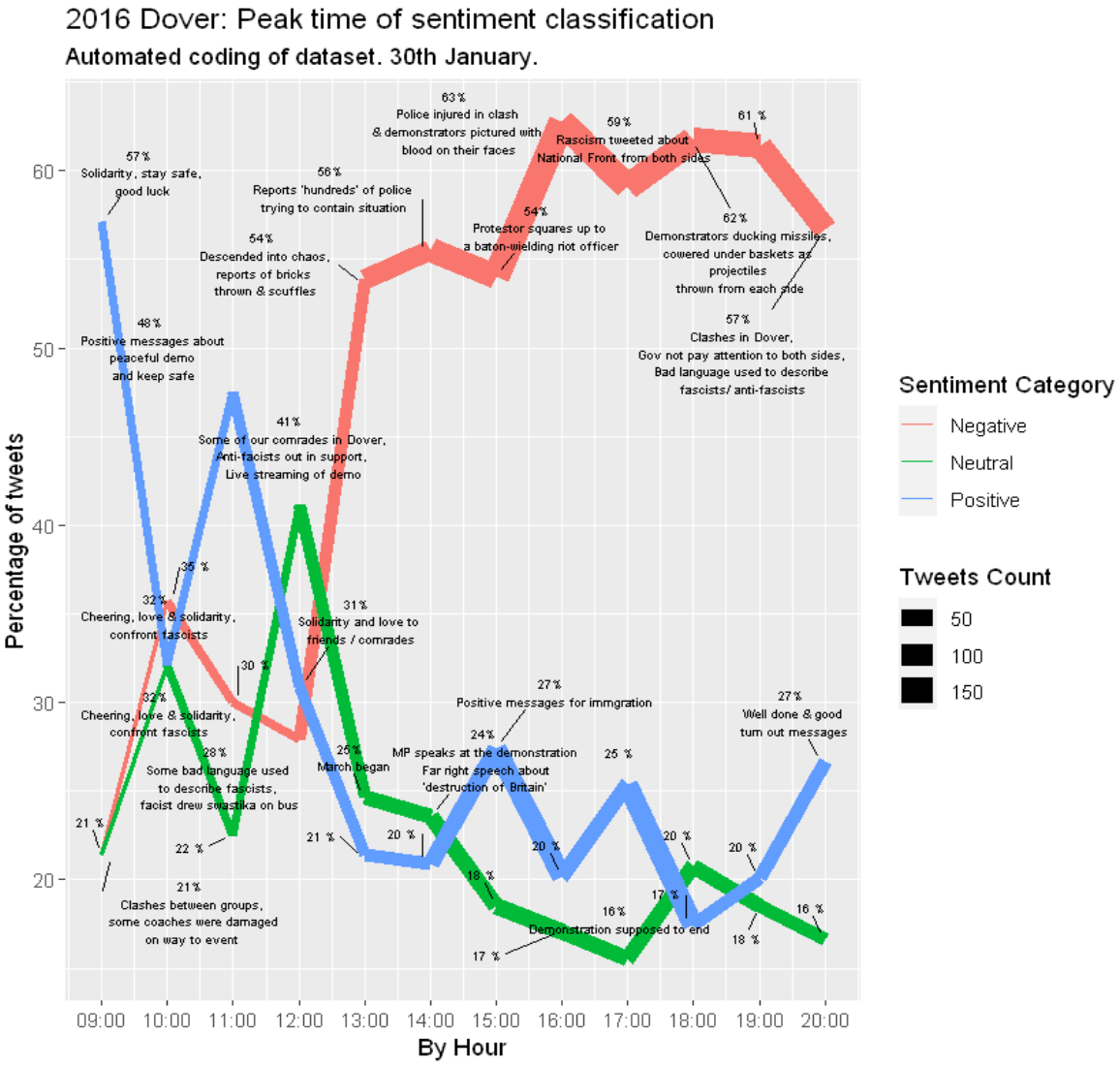

Figure 6.37 2016 Dover: Peak time of sentiment classification (automated)

In Figure 6.38, again shows another way of illustrating the sentiment where the red dots are individual tweet sentiment score and the black line is the average of tweets for a particular time slot on the day of the event between 14:00 to 19:00. The highest volume of tweets are displayed between 14:00 to 18:00. There was greater rise of

tweets towards 14:00, which kept building onwards to 18:00, then saw a decline later. The main bulk of tweets are negative compared to positivity, which is reflected in the average which seems to be the nearest hour and the scale on the horizontal axis only goes from 0.1 to -0.5.

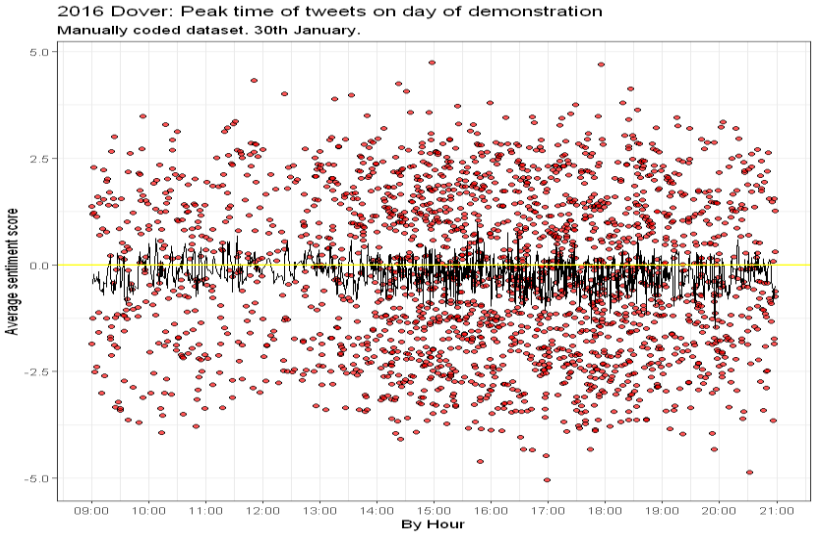

Figure 6.38 2016 Dover - Peak time of tweets on day of demonstration (Manual)

In Figure 6.39, the highest volume of tweets is between 14:00 and 19:00 which mostly agrees with the Figure 6.27, albeit with a slightly higher volume of tweets from the automated coded (relevant) data. The demonstration begun at 13:00 with 144 tweets, which saw a higher rise at 14:00 on 259. The count of tweets kept building at 15:00 to 280, dropped at 16:00 on 254, but peaked at 17:00 on 291, then saw a reduction towards 20:00 where it reduced by 159 to 132 tweets. In Figure 6.39 the average seems to be the nearest hour and the scale on the horizontal axis only goes from 0.1 to -0.9.

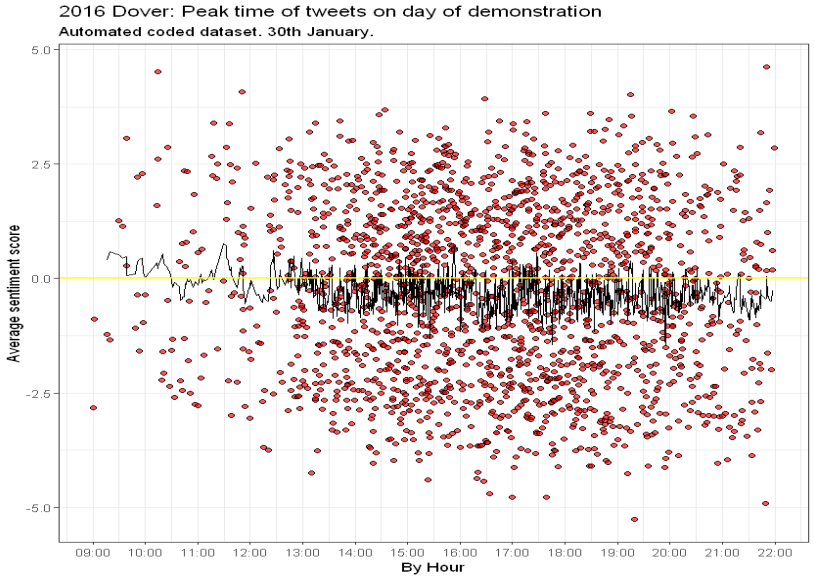

Figure 6.39 2016 Dover - Peak time of tweets on day of demonstration (automated)

Both Figure 6.38 and Figure 6.39 emphasised the volume of tweets and an average across time. However, Figure 6.40 provides a clear detailed view of the average score for the manual coded (relevant) data between 09:00 to 20:00, which mimics Figure

6.38. Figure 6.40 average score is mainly on the negative side ranging mainly from -0.03 to -0.31 except at 12:00 which is 0.02. At 09:00 is negative which grows in positivity until 12:00 on 0.02, but onwards the negativity grew to a peak of -0.31 at 16:00, then reduced from there to 20:00.

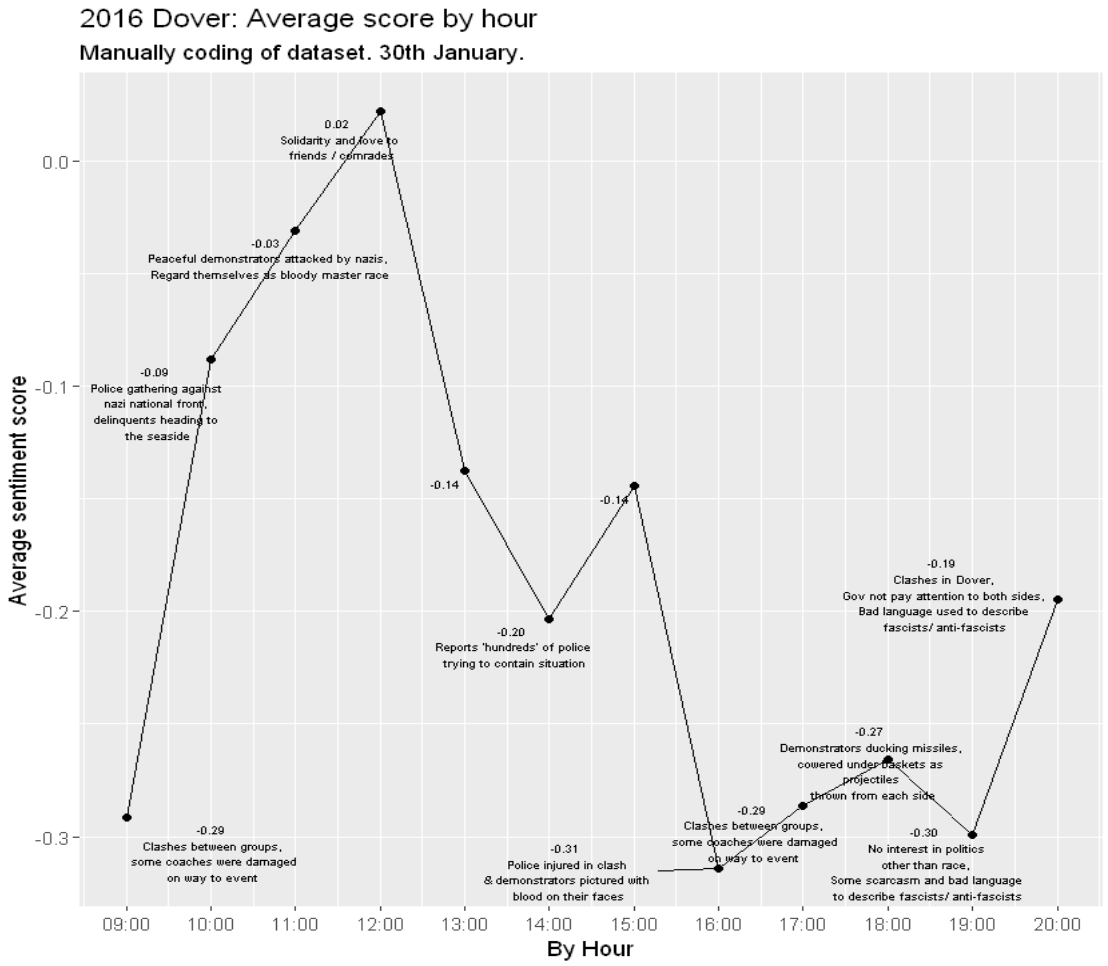

Figure 6.40 2016 Dover - Average score by hour overtime (manual)

In Figure 6.41 the average score is leans more on the positive side from 09:00 to 12:00 ranging from 0.01 to 0.32 with peak positivity at 09:00 on 0.32 except for 10:00 on -0.07. At 13:00 to 20:00 it is negative ranging from -0.26 to -0.45 with the peak at 16:00 on -0.45. Similar to Anti-Austerity, this demonstration positivity was higher before the event, but when it drew closer to the event negativity grew.

Figure 6.41 automated coded (relevant) data results compared with Figure 6.40 shows the average score follows a similar pattern, but automated is higher in negativity despite positivity being on a higher scale from the beginning. For instance, the most noticeable difference at 09:00 for Figure 6.40 where the average score is -0.29, but in Figure 6.41 it is the highest peak of positivity on 0.32, which highlights a 0.61 difference, which could be due to slightly more positive tweets in the few hundred extra tweets compared to total count of the manual coded (relevant) data.

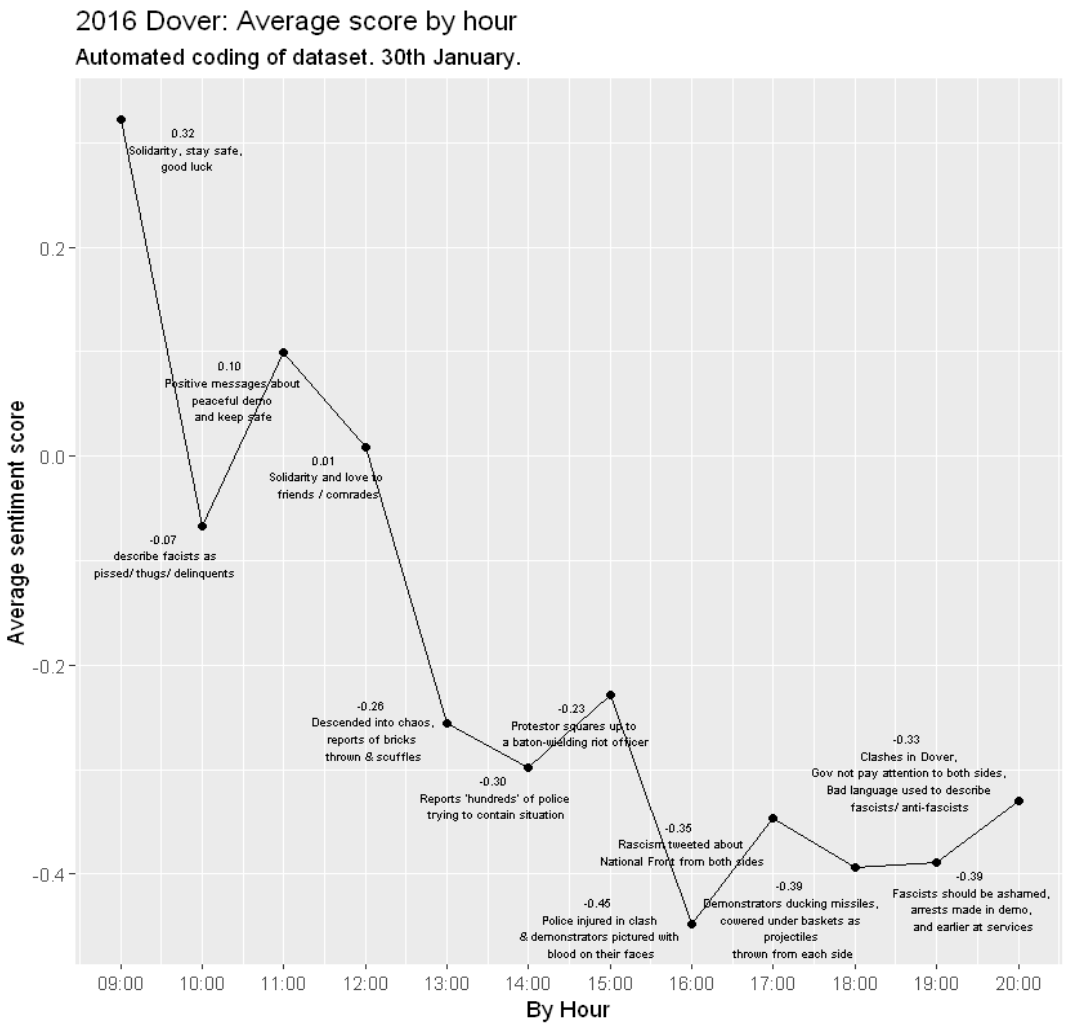

Figure 6.41 2016 Dover - Average score by hour overtime (automated)

We have explored the sentiment categories volume over time, average score and volume of tweets, which can help to determine the significant occurrences over time. In Figure 6.42, BinSeg is applied in the same way as before to identify changepoints for negative, neutral and positive categories for greater than 10 tweets which again has produced gaps in the period of time. The red line is limited to a maximum number of changepoints of 5.

In Figure 6.42 the negative line change points are from 13:00, 16:00, and 19:00 on the 30/01/16. At 13:00 the news reports (outlined in the above timeline) say chaos broke out with scuffles and bricks thrown, and at 16:00 coverage suggests police were injured in those clashes with demonstrators with blood on their faces. At 19:00 this coincides with reports that outline 'fascists should be ashamed and arrests made before the event and during it' (Couchman, 2018; BBC, 2016b; Gayle, 2016; Lennon, 2016; Nagesh, 2016b; Rkaina, 2016). The remaining change points on 31/01/16 are at 14:00 and 15:00. At 14:00 tweets outline 'annoyance that fascists getting blame for violence than other groups involved' and at 15:00 'cowering antifa and not nice person throwing bricks at nationalists'. The 01/02/16 after the event contain the lowest volume of tweets which makes it appear less significant. This could have been mis-

identified by BinSeg similarly to other events where negative line percentage is higher for a few tweets, but has more tweets than both neutral and positive at that same time, thus the proportion is greater which triggers a change point.

In Figure 6.42 the neutral line has three changepoints on 30/01/16, which are at 12:00, 13:00 and 15:00. At 12:00 both tweets and news reports similarly outline that 'our comrades operating with in Dover, antifascists begin the day, live streaming from the demo and looks like getting messy on the way to demo for some comrades' and at 13:00 'antifascists back on route, live stream from Dover and can you see press as nothing on tv'. Lastly, at 15:00 tweets suggest 'what is going on in Dover, follow updates on demo, what do you make of this, and more coming from us about what happened, and antifascists wear masks'. The remaining changepoints on 31/01/16 are at 10:00 and 17:00. At 10:00 the tweets outline 'part of the working class and protect themselves from fascists and is this what is meant by instant karma' and finally at 17:00 'anyone recorded allying with national front mates and left in bed with the elites'. The 31/01/16 after the event contain the lowest volume of tweets which makes it appear less significant and appears to follow the same issue outline for negative line and other events.

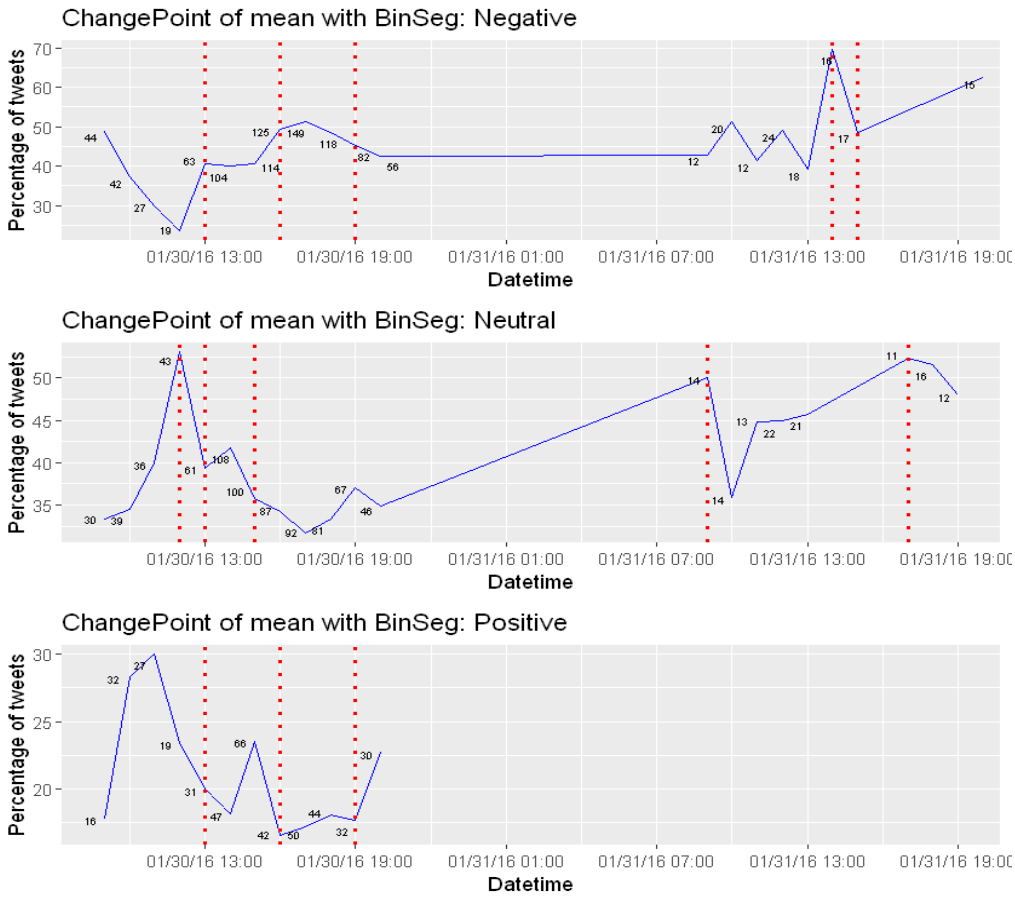

Figure 6.42 ChangePoint of mean with BinSeg by sentiment classification (manual)

In Figure 6.42 the positive line has the least volume of tweets for 30/01/16 and none for 31/01/16. There are only 3 changepoints on 30/01/16 at 13:00, 16:00 and 19:00. At 13:00 tweets outline 'antifascists stay strong, great solidarity and love, stay safe,

solidarity all standing up to fascists, and migrants are human beings in need of compassion' and at 16:00 'solidarity with all those fighting far right racist groups has no place, and well done to all the anti-fascist keep up the good work'. Lastly at 19:00 tweets suggest 'well done for showing your support, and well done for who stood up for British values'. Some of these tweets at these times are sarcastic/more negative than positive and belongs to the negative category.

In Figure 6.43 automated coded (relevant) data shows the negative lines change points are on 30/01/16 at both 15:00 and 23:00, and on 31/01/16 are 00:00, 10:00 and 11:00. The reasons behind the identification of these change points on 30/01/16 at 15:00 coincides with a news report outlines a 'demonstrator squared up to a riot officer', and tweets suggest at 23:00 'labour party traitors, anti-fascists are rich pretend to be poor and spend lives wanting no borders losers, and both are fascists there is no difference'. On 31/01/16 at 00:00, 10:00 and 11:00 the tweets outlined 'far right and anti-fascist protesters clash, bad language used to describe for and against demonstrators'. The changepoints identified for 31/01/16 after the event contain the lowest volume of tweets which makes it appear less significant. This could have been mis-identified by BinSeg similarly to the other 3 demonstrations and in the above manual coded (relevant) data results, where negative line percentage is higher for a few tweets, but has more tweets than both neutral and positive at that same time, thus the proportion is greater which triggers a change point.

In Figure 6.43 the neutral change points are only 15:00, 18:00 and 23:00 on 30/01/16 and throughout the event the number of tweets stayed pretty consistent level, which accumulated a total of 393 neutral tweets. This a very low number of tweets in comparison to negative and positive. At 15:00 the tweets suggest 'what is going on in Dover, what do you make of this, and more images of injuries/gets first aid for injury (evidenced in a news report which emphasised in the timeline of events at the beginning of this chapter) and being dispersed by police', and at both 18:00 and 23:00 the tweets further outline 'demonstration hardly mentioned in the news, and large demo marched from Victoria Square'. Many of the tweets are appeared to be incorrectly classified and should be categorised from neutral to negative.

In Figure 6.43 the positive change points on 30/01/16 are at 13:00, 15:00, 18:00, 20:00 and 22:00. This does not have the least volume of tweets this occasion but has 73 more tweets than neutral with a total of 466 tweets. However, there are a reasonably large proportion of the tweets for positive that need to be reclassified mainly for negative (many sarcastic tweets) than neutral. We have clearly identified there would be more tweets allocated to negative from neutral and positive through the observation of the tweets for both automated/ manual coded (relevant) data, which may be the significance of the results may be lessened. Despite concerns around classification of some data, at 13:00 the tweets outline 'stay safe, and love/solidarity all standing up to fascists', at 15:00 'respect to every one of you guys, not afraid to

stand up against fascists good on you and win against for anti-fascists congratulations to all' and at 18:00 'well played anti-fascists a few bleeding fascists, and well done/thank you for taking a stand'. Moreover, at 20:00 the tweets state 'karma my lovelies, and love/solidarity for all anti-fascists and stopping nazis is kind' and lastly by 22:00 'anti-fascist well prepared today and solidarity to brave anti-fascists, today worrying times, but good work'.

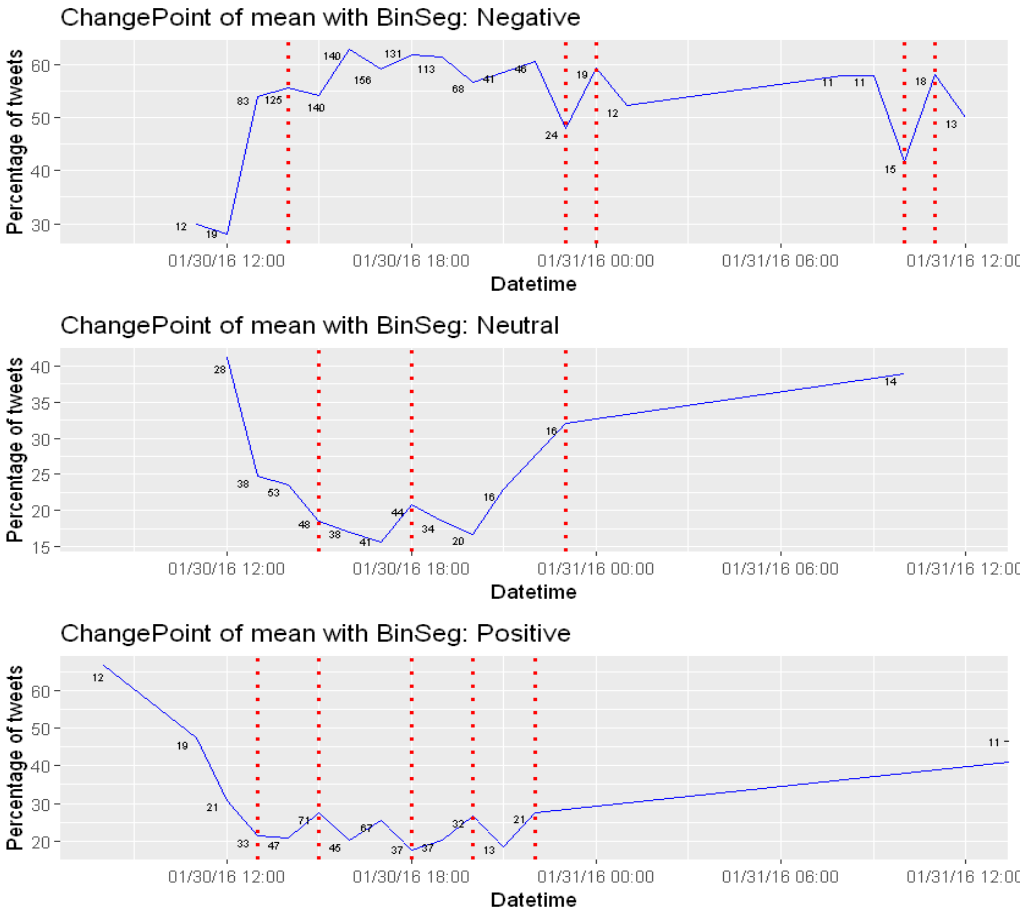

Figure 6.43 ChangePoint of mean with BinSeg by sentiment classification (automated)

[Intentionally Left Blank]

We can repeat the analyses for the other techniques used to score the sentiment. As an illustration the machine learning results of the predictions from both Naïve Bayes (NB) and Max Entropy (MaxEnt) are below. In Figure 6.44, the predictions from both Naïve Bayes (NB) and Max Entropy (MaxEnt) display majority for negative on the red line, which is practically 100% for NB with 0 for both neutral and positive categories, but for MaxEnt negative is approximately from 72% to 95% of tweets. Furthermore, Max Entropy does have a count for neutral between approximately 10% and 24% and positive roughly between 0% and 8% on a lower volume of tweets than neutral. There are no crossover points similar to 2016 MMM and Anti-Austerity events, but there is a mirror opposite with negative and neutral for MaxEnt, which is in common with the other 3 demonstrations.

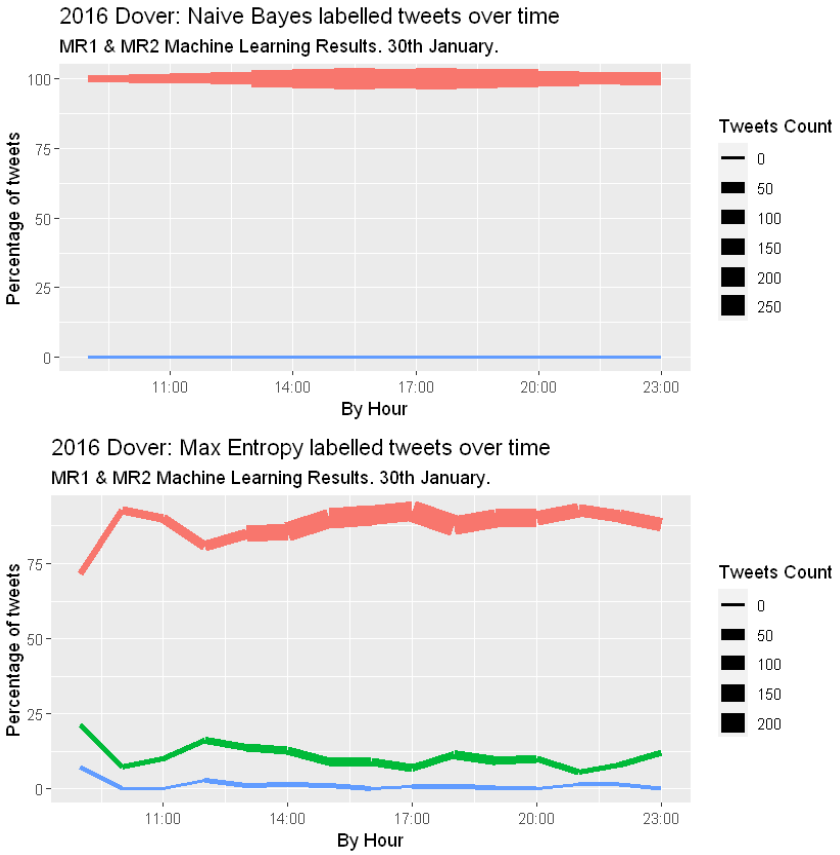

Figure 6.44 2016 Dover - prediction of sentiment by Naive Bayes/ Max Entropy over time (automated)

In Figure 6.44 both NB and Max Entropy result shares are limited in similarity to both Figure 6.36 and Figure 6.37 with negative being the majority with the highest volume of tweets. The part about the majority being negative shows that the predictions are somewhat supported by the initial analysis as in Figure 6.44. However, the thread of results has a limited connection with some of the peak/trough time of sentiment classification, average, and changepoint to the predicted results.

6.7.5 Summary

The results from both MMM events, Anti-Austerity, and Dover have indicated Max Entropy has a stronger connection with other results, but for Anti-Austerity NB is the strongest. However, both NB and Max Entropy have the strongest connection with other results, but both showed a weak connection in Dover results. Therefore, there is no definitive winner between each algorithm. The change points identified in each of the events on the day of the live demonstration has a connection in accordance with the news articles and tweets, which outlines the sparks of tension during the event. However, some of the changepoints appear inaccurate due to the very low volume of tweets at specific time intervals.

The BinSeg technique applied to identify significant change points was somewhat successful throughout the results for each demonstrations dataset. However, for all demonstrations results there were some changepoints that were questionable on their importance due to very low tweet counts. This might have been due to the issue previously outlined where one sentiment category percentage is higher in proportion for few tweets, but has more tweets than two other sentiment categories, thus the proportion is greater which triggers the change point. The way the graph is set and/or BinSeg settings could be enhanced to improve on BinSeg identification for change point(s) for the Twitter data. Furthermore, the low number of tweets might be more of the issue which could be further improved by both the data collection, pre-processing and coding of relevant data for both manual and automated processes of the data to increase the number of tweets which may balance out the graphs to provide more consistency to the timeline.

In future, further changepoint techniques could be explored into greater depth to analyse the data, such as R programming packages called “BCP” Bayesian Analysis of Change Point Problems, “ECP” Non-Parametric Multiple Change-Point Analysis of Multivariate Data, and “CPM” Sequential and Batch Change Detection Using Parametric and Nonparametric Methods which may be used in the analysis.

In section 7, we will evaluate the project strengths, weakness, improvements, and relate this back to the aim, research questions, objectives and deliverable. On the basis of the evaluation recommendations have been made for further research.

7 Evaluation and Recommendations

The evaluation will cover a wide range of points on what went well, not so well, and future improvements.

7.1 Pilot study

The pilot study (refer to both sections 4 and 4.1) was successful in helping to inform further study into the different demonstration cases, such as which tools and techniques to apply e.g., use of NVivo, R and its packages, Tableau and MongoDB. The researcher knowledge of the sentiment analysis is a new method, and much knowledge was gained in the process, such as understanding which R packages (e.g., 'dplyr' 'lubridate') are more effective in the cleansing, analysing, and visualising of the data, such as using word clouds to identify key words to describe topics of the event. In the research, it was discovered that automated method would be required to automatically label the data on a larger scale with use of machine learning and a way to evaluate the strength of the dictionary's outcome for future analysis, which led to precision, recall and f-measure. Alongside, the timeline of events by sentiment are helpful to understand the level of sentiment, but we identified change point techniques could make it clearer to determine why there are significant points of change at the event, in turn may be helpful for police to adapt for the event.

The pilot study analysis (refer to section 4.1) had limitations due to the gaps in the dataset's timeline of events due to the limitations of the Twitter API access and functionality of NVivo that was manual process rather than an automated with a constant stream of data (refer to section 4.1). The proportion of emotion by multiple categories has more tweets identified as 'Null' as could not be assigned to another category and when categories by polarity, then there is misclassification to extent as some tweets identified neutral was negative/ positive. These results helped to further understand that 'Berkeley' dictionary does not perform well on this dataset and to investigate the dictionaries more closely along with other evaluation techniques to help determine the strength of the outcome (refer to section 4.1). The incompleteness in the data meant that an alternative approach to acquire the dataset from Twitter, which led to use of DiscoverText to buy the tweets. Additionally, the data is collected from the USA demonstration, but the cases applied in the main study for the project is applied into a UK context, so some tools and techniques may have been less helpful for future analysis. In future study, it would have been interesting to test the American dictionaries on American tweets and then compare against UK setting for their F1 scores in a public order context.

7.2 Initial Data and Information Processing

The initial findings for each case study (refer to section 5) helped to understand the context of the demonstration and graphs produced by DiscoverText provided a quick insight into the data over time. The volume of tweets overtime could be filtered by day and a detailed view by day and hour through time intervals. This helped to peaks and troughs of events before, during and after the event and to compare each of the cases. Furthermore, there was auto-generation of tables, such as top hashtags provided insight on top potential topics related to the event, such as Anti-Austerity highlighted #4Demands which is organisation designated one for the event and different hashtags related to other topical news, such as #TFL and #Paris. The word clouds that were useful in the pilot study helped to shape an overall image of the event (refer to section 4.1), which was re-enforced with other insights, such as change point analysis results. The lexical diversity when analysing the tweets helped to establish a low diversity of words used, but this is dictated to an extent by Twitter character limitation of 140 characters, but this may be different now with the higher character range of 280. However, Term Frequency (TF) and Term Frequency Inverse Frequency (TF-IDF) are helpful to discover prominent keywords used in the event, and words that should have been included in the stop-words list (refer to section 5), so this helped to remove words to even the scale of importance in the outcome of the analysis. Zipf law has been identified to conduct in future work for the dynamic generation of stop words list, as the results were impacted by the removal of words e.g., 'not' and 'wouldn't' from the pre-compiled list, which could have made slight difference to the sentiment outcome (Saif et al., 2014).

The extraction of data based on the key terms to retrieve the most relevant data based on each event provided the data for the analysis. There was relevant data for each event, but it how the data was searched for could have been improved with terms used and use of its search capability narrow down the area to ensure greater number of tweets are topics discussed for the event than being irrelevant area. For Anti-Austerity, the term 'London' is too generic that showed a large proportion of irrelevant tweets before, during and after the event (refer to both sections 4 and 4.1). The free tools and techniques applied to identify the terms with higher relevance for the event, needs to be improved with alternative methods to be explored. The transformation of the data to remove stop words, numbers and special characters was successful in a way that it cleans the dataset to prepare the data for analysis. There were problems uncovered overtime in this process, which are: -

- The hashtags were removed as it appeared most did not contribute to the overall sentence. However, overtime it was observed that there were more hashtags used as part of the sentence, which may have slightly impacted the sentiment output for some tweets. In future, it would be wise to test the sentiment output with and

without hashtags to assess the impact as which pathway would strengthen the outcome.

- Issues with file versioning have caused at most minor problems in the process, which in part lacked documentation. This led much time being spent to investigate the changes from one version to the next one. The main problem caused was when the date/time was different for two datasets, which was not identified to late on when the columns were split for date and time. The date was correct, but the time was inaccurate due to not identifying it had text of AM and PM when the time was split to take that into account, whereas other two were 24-hour clock time than 12-hour one.

- The use of Excel spreadsheet to calculate the precision, recall and f1 score, etc could have been place into a database to query and calculated with a written R program to automatically calculate it than the manual process which makes it non-repeatable.

The extraction and transformation have strong elements to the process that worked well and not so well as outlined above. The learning from this help benefit the researcher and potential new researchers to the area to avoid the pitfalls of the project and gain insight on what is required sentiment analysis to enhance the process ensure a higher quality output.

7.3 Ethical and Legal Complications

This section of the chapter has been published in the conference paper (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018), which is provided in appendix 10.9 and from which we draw from throughout.

Datasets with this scale of social interaction, speed of generation and level of access are unprecedented in the social sciences way (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018). Therefore, some universities may have not caught up with the pace of technology and this is often reflected in their ethical policies, so there may be ethics panels have already scrutinised such data, they may still deem it to be ‘public data’ due to the lack of a suitable framework to evaluate the potential harm faced by those whose ostensibly public data is used way (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018). In some cases, ethical approval is not required per se, but it is suggested by a given university's policy that researchers consult resources, such as the Association of Internet Researchers (AoIR), that can help to ensure that any social media data are used in an ethical fashion way (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018). Additionally, social media platforms require a user to adhere their terms and conditions to use and share data, for instance, a user can only share the Twitter ID from the data collected which can pose issues of reproducibility of research and the protection of the social media users.

Despite noting above that some ethical panels are not making much consideration about the ethical use of social media data, there is some evidence to suggest that a number of universities are making strides towards updating their ethical guidelines with regards to social media data (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018). As one such example, the University of Sheffield has a research ethics policy note that raises many important points that can be considered in other institutions (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018). This note indicates that research must have ethical approval before a dataset can be extracted. However, this may pose both a financial and a contemporaneity problem. If the researcher wants to use historical data that will in any case come at a cost then this will be the case with or without prior ethical authorisation (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018). However, if the data cannot be extracted on-the-fly because ethical approval is taking time to obtain, then the institution's budget would have to be prepared to pay for those data in the long term. Furthermore, if the researcher is considering topics of current interest and wishes to amend their search criteria as data come in, it may not in fact even be possible to seek suitable a priori approval (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018). Of course, planning in advance is well advised here, but there are times when one cannot predict the topics of research interest that will arise today, tomorrow or in many weeks' time, which makes it difficult to plan such requests in advance (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018). This policy is thought provoking, as it makes the researcher think about the importance of ethics in the very early stages of their research and the requirement for ethical approval for social media research is clearly a step in the right direction towards ensuring high ethical standards (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018). However, as noted above, it may be financial unviable, or prevent the collection of data required for some projects. To that end, we would recommend that perhaps there be a fast-track ethical approval system for time-critical social media data projects so that on the one hand they receive suitable ethical scrutiny, while on the other they can also proceed in a timely manner, enabling researchers to react to current events of public interest (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018).

The concerns raised above, demonstrate social media research ethics is an area that requires further development and awareness to ensure the public's data are represented in a context that is respectful, accurate and in a fair way (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018). This topic is a hotly debated area of research in the community (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018). The application of ethics must consider the concerns raised above. If researchers and organisations are not careful in their approach, the disconnect between researchers and users may grow further. A lack of action regarding such ethics could lead to a series of undesirable consequences, such as users calling on social media platforms for changes in their terms of service to restrict the use of their data. The impact of this may make it extremely difficult to use social media data for research designed for the public good.

In the project, Twitter terms and conditions are followed, alongside this the Association of Internet Researchers (AoIR) was provided as a guideline to follow, and at the time of consulting this resource, there were no specific procedures to follow that were specific to social media data way (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018). Therefore, other resources provided by The Economic and Social Research Council (ESRC) Social Media Stewardship (SMDS), IPSOS MORI and The Government Social Research (GSR) were reviewed (refer to section 3.1), where some contained a few or more guidelines to follow. In accordance with the guidelines, we kept most of the tweets used in the project have not been publicised. If the researcher needed to provide information based on a tweet or a series of tweets, then largely the tweets have been summarised and/ or been re-written, so the Twitter users remain anonymous.

Both concerns in the project and as a whole outlined above regarding the ethical challenges of using social media data can make for a difficult challenge for the social media researcher (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018). The best course of action the researcher community can take is to address concerns and difficulties on case-by-case basis, thereafter trying to update guidelines and frameworks to deal with such cases. Genuine mistakes might have been made in the research community, which both individual researchers and the community as a whole can learn from. If a researcher has made a genuine ethics-related mistake in their work and has demonstrated remorse, then we as a community need to forgive and look to further strengthen the ethical standards and frameworks available to us. Indeed, ethical concepts are not just hoops to jump through in the early phases of research, but concepts requiring ethical inquiry (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018), which may in itself take time. Mistakes may not be recognised until well after they have occurred and numerous judgements are possible, which can provide uncertainty and ambiguity, but this is likely to apply to any research (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018). Ethical considerations will be in a constant state of assessment throughout any project and each case that arises during the research process can be worked through using a set of context-specific decisions. In addition to this, researchers must be guided by core ethical principles set by their employing organisations and external bodies, while also employing an appropriate mixture of the frameworks as laid out above, to ensure that the highest ethical standards are followed in any research.

In summary, there is a need to improve ethical assessment and one way to do this is to create a value-based ethical culture and practices in the research community and within other organisations for the development and deployment of intelligent systems both within the UK and elsewhere. This is known as Value Based Design (VBD) (Baldwin, Brunsdon, Gaudoin & Hirsch, 2018). To do this, one must identify, enhance and ultimately embrace management strategies and social processes that facilitate value-based ethics within their design process. This could be included as an additional

step in the framework in a future development, as it may provide a way to ensure a higher standard of ethical practice in the future.

7.4 Dictionary approach and Machine Learning approach

The dictionary approach produced the strongest F1 score out than any other approach, such as machine learning and gold standard sentiment outcome. The strongest dictionaries are Jockers family, Bing, Loughran MacDonald and SentiStrength (refer to both sections 6.5.2 and 6.5.3). The less strong dictionaries for this topic area are Berkeley, Socal Google and Senticnet. However, Berkeley was used for pilot study and tended to identify the positive category more often (refer to section 4.1) and in the case studies showed a similar result with neutral and negative further behind on occasions (refer to both sections 6.5.2 and 6.5.3). Additionally, the pilot study showed Berkeley incorrectly categorised some tweets which again remains the same and on a general level has a higher misclassification rate as proven in its outcome, which is why it's one of the weakest outcomes (refer to section 4.1). The introduction of new categories on top of the polarity of negative, neutral and positive along with somewhat negative, neutral and positive produced weaker outcomes (refer to appendix 10.12.4). This could be argued further with the pilot study where it showed a much weaker F1 score with multiple categories of emotion, as it may be more difficult to define which aligns with the correct sentiment category leading to higher levels of misclassification (refer to section 4.1). It appears from the results that a smaller number of categories leads to a stronger output, but misclassification remains a prominent issue in the field due to other factors, such as use of sarcasm being difficult to detect in what is expressed in the opinion (refer to section 5.8).

The dictionary approach, machine learning and gold standard phases requirement improvement to help strengthen the outcome and some suggestions above may help with that, such as Zipf Law. In further research, there are other considerations made to each stage of the process, which are as follows: -

- The manual classification of tweets classified needs to be increased for greater generalisation. Also, an increase manual classifier could have helped to further check the reliability of the process, as the result showed greater agreement between MR1 and MR2, but MR3 (refer to sections 6.4 to 6.4.5) on a much lesser scale. If the more tweets are manually classified, then this may lead to a larger training and test dataset. This may help to provide greater balance in the sample, as there was class imbalance when it can to the training sample as there were more negative and neutral tweets over positives ones (refer to both sections 6.5.4.3 and 6.6.1). If over-sample the minority class, and/or under-sample the majority class to reduce the class imbalance to provide clearer detection of positive tweets.

- An evaluation step could have been conducted on the automated coding of the relevant/irrelevant data score with the form of recall, precision and F1. The percentage of relevant tweets (refer to section 6.2) found that the recall will be low, but the evaluation techniques could have helped to understand whether the precision is low or high as well. The evaluation could have helped to determine if the tweets more or least reliably relevant which would help improve analysis.

- Word2vec (Al-Saqqa & Awajan, 2019; Alshari, Doraisamy, Mustapha & Alkeshr, 2017; Lamaute, Luo, Finkelstein & Cotoranu, 2017; Liu, 2017) is a method of Bags of Words and TF-IDF that do not encapsulate the meaning between each of the words and focus on the separate words as features. Whereas, Word2Vec are words mapped to number vectors that sum up the semantic meaning of the words.
    - Word embeddings use a model to put a word in a vector that similar words are closer to each other, such a negative word that are adjectives will be close to one another, therefore, encapsulates syntactical and semantical information of words (Al-Saqqa & Awajan, 2019; Alshari, Doraisamy, Mustapha & Alkeshr, 2017; Lamaute, Luo, Finkelstein & Cotoranu, 2017; Liu, 2017).
    - There are multiple ways (Al-Saqqa & Awajan, 2019; Alshari, Doraisamy, Mustapha & Alkeshr, 2017; Lamaute, Luo, Finkelstein & Cotoranu, 2017; Liu, 2017) of word embedding vectors, and a way is to use Word2Vec that uses neural model to learn, which has to architectures that can be applied which are 'Skip Gram' (predicts embeddings for the surrounding context words in the specific window given a current word) and 'Continuous Bag of Words' (predicts the word under consideration given context words within specific window).

- The process adopted to convert text into a vector was performance intensive as it would create it as a single vector, but 'text2vec' constructs the document term matrix in a memory friendly way because the package is written in C++ which makes it efficient, so users do not have to load all data into RAM (Selivanov, Bickel and Wang, 2020; Selivanov et al., 2020). The cross-validation process would not complete its job on some occasions due to a resources error. Therefore, there was no result for many of the outcomes, which is why this was not considered in the analysis.

- The strongest machine learning algorithms are Naïve Bayes and Max Entropy for the case studies (refer to both sections 6.5.5 and 6.6.1.2). However, the F1 score may increase if the configuration of the algorithms hyper parameters are changed from the default settings to enhance the outcome.

- The Gold Standard applied produced a reasonable outcome despite being lower than dictionary approach and other ways could be explored to improve the F1 score. Additionally, it would have been a fairer test if Naïve Bayes algorithm were used to compare against dictionary approach. The Naïve Bayes encapsulated in the 'Caret' R package was not applied due to limited knowledge on how to apply it to tweets.

- The combined dictionary performed reasonably well throughout each of the approaches, and the cut-off point established helped to balance out the combined dictionary between the sentiment categories (refer to section 6.5.1). However, it may be that further enhancements could be made to improve it, which are as follows: -
    - Combined dictionary had the largest list of words for sentiment, but performed less well compared with ones with much less words. There could be further work in refining the importance of the sentiment weight of a word, words list could be refined to be more tailored for public order events and need to provide balance with the number of words for negative or positive or neutral in the list.

- The majority voting category for the dictionary approach list of dictionaries included to vote could have been reconsidered to remove few of them to increase the accuracies of the decision on which sentiment category has the majority (refer to section 6.5.1.1). Additionally, other dictionaries that are new in the future could be experimented with to identify if it could enhance the process. These changes could reduce the need for a coin flip to decide on the majority sentiment category for a tweet. Furthermore, the coin flip function proved useful, but alternative ways could be explored to identify if there is any way to improve the outcome on how the deciding vote is cast.

The dictionary, machine learning and Gold Standard approaches have performed well in the 0.60s for some of the results, but dictionary outperformed compared to the other approaches. As previously identified above further improvements can be made in the future, which may enhance the strength of outcome. These changes could help each approach more or less in some ways on the quality of output.

7.5 Change Point Analysis Approach

The change point analysis helps to identify significant points of change before, during and after for each event for manual classified and automated classified based on public order keywords, which many of these points are supported by the news media and tweets on Twitter (refer to section 5.11). This was shown in the many graphical images that looked at peak time of the event day based on sentiment, average sentiment score and use BinSeg to identify significant change points (refer to section 6.7).

BinSeg was used for change point, but other techniques were considered, such as AMOC and SegNeigh which showed no significant difference for the specific datasets to further understand the event (refer to section 5.11). BinSeg helped to identify many significant change points, but some were deemed inaccurate due to low volume, how the correct percentage of proportion is high for a low volume of tweet may trigger a change point and the weak link between news media and tweets (refer to section

6.7.5). There is a need for greater investigation into other methods of changepoint and to understand how BinSeg is allocated to identify the reasons why some are picked that do not seem to hold much significance in its change.

The evaluation has covered a wide range of the process from pilot study to the change point analysis. The next step is to produce recommendations based on the evaluation and assessment of the deliverable.

7.6 Review of Objectives

The project aimed to analyse social media in the context of public order events. There are a series of objectives to help achieve the aim, which can be referred to in section 1.2. Below, explores the objectives to identify if they have been met, which are as follows: -

- **Objective 1 & 2:** The literature review has provided information on the historical and ongoing development of police practice usage of social and technologies used in managing public order. Additionally, the application of sentiment analysis different approaches was evaluated in a police context in a limited way, but mainly in others, such as reviews for a product or restaurants.

- **Objective 3:** A series of social media platforms are considered in the project, but the most applicable is Twitter due to the openness of the platforms and it tends to be used more to post public messages on public order events as emphasised in both the literature review and social media research strategy sections.

- **Objective 4:** A suitable range of methods and instruments were investigated. As a result, a social media lifecycle was identified on the UK Government social research team, which is adapted for research than commercial use. This was applied to the pilot study and each of the UK demonstration cases.

- **Objective 5 & 6:** The Million Mask March (MMM) 2015 & 2016, Anti-Austerity 2016 and Dover 2016 were chosen as cases to study. The data mining and text mining techniques were explored to extract, pre-process, and analyse the data, such as R to prepare, explore and analyse the data and DiscoverText to extract the datasets. The keywords needed to be identified for each event to ensure the most relevant data is extracted from Twitter API. These keywords were identified with the use of Twitter advanced search, news media and hashtagifme, which are used within DiscoverText to extract the data. Numerous R packages were identified in the pilot study and outside of it to help prepare and analyse the data, such as ‘TM’, ‘GGPLOT2’, ‘Lubridate’, ‘RTextTools’, ‘Caret’ and ‘DPLYR’.

- **Objective 7:** A relevant methodology has been adapted and appropriate evaluation has been conducted to identify the strengths and weaknesses of the project, which a series of recommendations have been made below on how to enhance future sentiment analysis projects in the realm of social media.

The objectives (refer to section 1.2) helped to achieve the aim of the project and answer a few research questions (refer to section 1.1), which will now be explored as to whether they have been met: -

- The first question has been achieved as significant points of change have been identified based on sentiment which has been effective in aligning with current events at the physical event.

- The second question has been achieved as a series of dictionaries have been identified that could be used again in public order events that could determine potentially a higher level of accuracy.

- The third question has shown that dictionary approach is more effective F1 scores than machine learning approach, however, there is improvement required in the framework produced in this project.

The objectives and research questions have been evaluated and have shown to meet the aim of the project and the deliverable. The recommendations based on the evaluation will be explored in the final chapter.

7.7 Recommendations

There are series of recommendations for improvement based on the evaluation strengths and weaknesses for future work, which are as follows: -

1. The project established that a dictionary approach produced a stronger outcome, and that a use of machine learning process for hybrid approach helped to identify the sentiment classification on a greater scale (refer to section 6.6.1.2). Some of the key dictionaries identified in this project includes Jockers family and SentiStrength (refer to both section 6.5.3 and 6.5.4). Additionally, the algorithms that produced the strongest outcome are naïve bayes and maximum entropy (refer to section 6.6.1.2). However, these were key in this project, and may be for other public orders events and other topic areas.

2. To improve the combined dictionary results (refer to section 6.5.4) there is a need to ensure UK English terms are applied in the lexicon, identify whether any other UK dictionaries are developed that could be used to combine with it. Moreover, remove any dictionaries that performed less well and perhaps re-scale the

sentiment score of an appropriate proportion in a UK context in the combined dictionary.

3. To increase the generalisability of the trained set to balance it out for negative/ neutral and especially positive category (refer to the summary of results in both section 6.5.4 and 6.6) as this was a consistent issue in the project.

4. In stop word removal use an alternative approach to ensure that less keywords get removed (refer to sections 3.1.6.2, 4.1 and 5.12.2) that are pertinent for sentiment analysis process. A key suggestion is to apply Zipf Law (refer to section **Error! Reference source not found.**) to draw up the list of keywords as it has been effective in other projects.

5. To increase the keywords list for public order to filter relevant and irrelevant data for coding in an automated way. Some of the results especially before or after the event are more irrelevant (refer to the change point summary of results in section 6.7.5), so fine tuning the list may have helped to reduce this issue, and make for a more relevant tweets to be included in the study. This may have led to different change points, peaks and troughs for other graphs.

6. To produce precision, recall and F1 scores for the automated coding of relevant/irrelevant to determine if the tweets more or least reliably relevant which would help improve analysis (refer to section 7.4).

7. The change point analysis requires further examination of alternative techniques to explore the dataset(s) to identify if any improvements can be made in the future (refer to section 6.7.5). Additionally, to understand how BinSeg allocates a significant point of change to understand reasons for the issues outlined in the evaluation (refer to section 6.7.5).

8. To improve ethical assessment with the use of Value Based Design (VBD) to create a value-based ethical culture and practices in the research community and within other organisations for the development and deployment of intelligent systems both within the UK and elsewhere (refer to section 7.3). This could be included as an additional step in the framework in a future development, as it may provide a way to ensure a higher standard of ethical practice in the future.

A set of recommendations produced can be applied into future project that further enhance the sentiment analysis projects for research and co-commercial use. A conclusion will be drawn in section 8, with the main findings of the project.

8 Conclusion

In conclusion, the project has had a main series of successes, limitations and improvements that are required in the future, which are as follow:

- The datasets collected are limited to four UK demonstrations that are only sourced from Twitter. The datasets extracted are based on a series of keywords identified for each of the events. The was a small proportion of tweets only focused on for the sentiment analysis approach. Some of the data collected was relevant, but there was a fairly high level of irrelevant or of poor-quality data (refer to both section 6.2 and 6.2.1). Further improvement could have been made on the keywords chosen for extraction of the data and also an enhanced list for relevant/ irrelevant keywords for public order event(s) to classify data into their correct category with an evaluation of reliability based on precision, recall and F1 (refer to section 7.4). This will help to increase this percentage of relevant to a higher number further work is required to increase keywords for both relevant and irrelevant to accurately identify the tweets.

- The combined dictionary was created to combine a series of the dictionaries and standardised, of which the results were somewhat in the middle of the 19 dictionaries in terms of performance, so improvement have been specified in point 2 of the recommendations (refer to section 7.7). This may be the one of few dictionary's where multiple dictionaries have been combined, but is evident due to the limitations identified in this project (refer to section 6.5.1).

- The dictionaries results of breakdown showed some of the dictionaries outperformed against other F1 results mainly with highs up in the late 0.60s (refer to recommendation 1 in section 7.7). This was further emphasised their strength in the machine learning results with algorithms, such as Naïve Bayes and Maximum Entropy mainly in late 0.60s and few in early 0.70s (refer to recommendation 1 in section 7.7). There were some notable changes in the framework, which may have made a positive difference to the output, such as different technique for removal of stop words (refer to recommendation 4 in section 7.7). The project used a large number of dictionaries compared with most publications review, which provides a greater insight on the successes and limitations of dictionaries.

- The machine learning results with the use of tweets and gold standard applied showed strength in their results, but less strong compared to dictionaries results as few algorithms more situated in lower 0.60s (refer to section 7.4 and recommendation 3 in section 7.7). This showed dictionaries performed better, however, only the default algorithm settings were applied, so further improvement

on the framework to enhance the F1 scores, such as tweaking of algorithms default settings and generalisation of the sentiment categories (refer to section 7.4).

- The change point analysis helped to identify some significant changes points, which solidified the initial results in the background of each event. There was much correlation between the significant change points and the news media timeline, however, on occasion this was supported by many tweets (refer to section 7.5 and recommendation 7 in section 7.7). However, few tweets were identified as a change point, which are incorrectly stated in the graphs to irrelevance of tweets and too few tweets certain times. The basics of change point analysis with sentiment based data was applied, and limited to use of BinSeg technique, so further research in this area could help gain greater insights into the data (refer to section 7.5).

- The social media lifecycle adapted for research that was tested, but will require further experiment with new projects to identify any further improvements (refer to both sections 3 and 3.1). The ethical consideration in the project showed that the ethical framework for social media data requires further development to form a standardised approach, as the current position is somewhat confusing for a beginner to grasp as the guidance on the use of the data could be made clearer for technical projects to understand which stance to apply in a social media project (refer to section 7.3). Therefore, training and guidance at earlier stages would be helpful for a beginner to move forward with their research, and in time with further developments of standardised framework for ethics can be developed upon in the research community (refer to recommendation 8 in section 7.7).

As previously outlined above, the study main points highlighted the key benefits, limitations and areas for future work in the realm of text mining, sentiment analysis and ethics in a social media context for public order events. The sentiment analysis framework for social media requires further development in the research field, which can be known as social media data mining. The project notably came across short text issues with text pre-processing methods and how dictionaries/ algorithms face issues. There are key evaluations and recommendations that have been indicated to in section 7.

9 References

Acosta, J., Lamaute, N., Luo, M., Finkelstein, E., & Andreea, C. (2017). Sentiment analysis of twitter messages using word2vec. *Proceedings of Student-Faculty Research Day, CSIS, Pace University*, *7*, 1-7.

Adam, E.C. (1993). Fighter cockpits of the future. In: *Digital avionics systems conference, 1993. 12th DASC., AIAA/IEEE,* 318-323.

Agarwal, N., Lim, M., and Wigand, R. (2014). *Online Collective Action: Dynamics of the Crowd in Social Media.* Dordrecht, Springer Vienna.

Aggarwal, C.C. (2015). *Data mining: The textbook*. Cham: Springer.

Aggarwal, C.C., and Reddy, C.K. (2014). *Data clustering: Algorithms and applications*. Boca Raton: CRC Press.

Aggarwal, C.C. (2011). *Social Network Data Analytics*. Publisher: Springer US. ISBN: 9781441984616.

Ahmed, W. (2017). Ethical Challenges of Using Social Media Data In Research [online]. Last accessed 11th December 2014 at: https://www.youtube.com/watch?v=VeFMqL4Hj60

Akhgar, B., and Staniforth, A. (2014). *Cyber Crime and Cyber Terrorism Investigator's Handbook*. 1st Edition. Syngress.

Al-Saqqa, S., & Awajan, A. (2019). The use of word2vec model in sentiment analysis: A survey. In *Proceedings of the 2019 International Conference on Artificial Intelligence, Robotics and Control* (pp. 39-43).
Alshari, E. M., Azman, A., Doraisamy, S., Mustapha, N., & Alkeshr, M. (2017). Improvement of sentiment analysis based on clustering of Word2Vec features. In *2017 28th international workshop on database and expert systems applications (DEXA)* (pp. 123-126). IEEE.

Amin, A. (2003). Unruly strangers? The 2001 urban riots in Britain. *International journal of urban and regional research,* **27** (2), 460-463.

Aminikhanghahi, S., and Cook, D. (2016). A survey of methods for time series change point detection. *Knowledge And Information Systems*, *51*(2), 339-367. doi: 10.1007/s10115-016-0987-z

Armstrong, C. (2017). R00ting the Ingroup: Anonymous and Social Identities in the Digital Sector. Doctor of International Conflict Management Dissertations. 15.

Association of Chief Police Officers (ACPO). (2015). Association of Chief Police Officers Submission to the Police Remuneration Review Body [online]. Last accessed 12th August 2015 at:

http://www.npcc.police.uk/documents/reports/ACPO%20submission%20to%20PRRB%202015%20Final.pdf

Association of Chief Police Officers (ACPO). (2012). The New Policing Landscape - the Role of the Association of Chief Police Officers [online]. Last accessed 24th August 2015 at: http://www.npcc.police.uk/documents/FoI%20publication/Disclosure%20Logs/Presidential%20FOI/2014/0018%2014%20Att%2002%20of%203%20New%20Policing%20Landscape%20role%20of%20ACPO.pdf

Association of Chief Police Officers (ACPO). (2010). Manual of Guidance on Keeping the Peace [online]. Last accessed 11th December 2014 at: http://www.acpo.police.uk/documents/uniformed/2010/201010UNKTP01.pdf

Athar, A. (2014). Sentiment analysis of scientific citations. [online] Cl.cam.ac.uk. Last accessed 28th March 2019 at: https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-856.pdf

Azevedo, A., and Santos, M. (2008). KDD, SEMMA AND CRISP-DM: A Parallel Overview [online]. Last accessed 12th August 2015 at: http://dis.unal.edu.co/profesores/eleon/cursos/md_2014/documentos/metodologias.pdf

Babuta, A., and Oswald, M. (2018). Machine Learning Algorithms and Police Decision-Making - Legal, Ethical and Regulatory Challenges [online]. Last accessed 5th August 2021: https://static.rusi.org/201809_whr_3-18_machine_learning_algorithms.pdf.pdf

Babuta, A. (2017). Big Data and Policing: An Assessment of Law Enforcement Requirements, Expectations and Priorities [online]. Last accessed 15th April 2019 at: https://static.rusi.org/201709_rusi_big_data_and_policing_babuta_web.pdf

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec* (Vol. 10, No. 2010, pp. 2200-2204).

Baker, S.A. (2012). From the criminal crowd to the "mediated crowd": the impact of social media on the 2011 English riots. Safer Communities, Vol. 11(1) pp. 40 - 49. Last accessed 12th August 2015 at: doi: http://dx.doi.org/10.1108/17578041211200100

Baldwin, J., Brunsdon, T., Gaudoin, J., and Hirsch, L. (2018). Towards a social media research methodology: Defining approaches and ethical concerns. *International journal on advances in life sciences*, **10**.

Bali, R., and Sarkar, D. (2016). *R machine learning by example*. Birmingham: Packt Publishing Limited.

Bailo, F., and Vromen, A. (2017). Hybrid social and news media protest events: from #MarchinMarch to #BusttheBudget in Australia. *Information, Communication & Society*. 20 (11), 1660–1679.

Barisione, M., and Michailidou, A. (2017). *Social media and European politics: Rethinking power and legitimacy in the digital era*. London and Basingstoke: Palgrave Macmillan

Bartlett, J., and Norrie, R. (2015). immigration on twitter: understanding public attitudes [online]. Last accessed 12th August 2015 at: https://www.demos.co.uk/files/immigration_on_twitter.pdf?1428506056

Bastos, M., Recuero, R., and Zago, G. (2014). Taking tweets to the streets: A spatial analysis of the Vinegar Protests in Brazil [online]. Last accessed 12th August 2015 at: http://journals.uic.edu/ojs/index.php/fm/article/view/5227/3843

Batrinca, B., and Treleaven, P. (2015). Social media analytics: a survey of techniques, tools and platforms. *Ai & society,* **30** (1), 89-116.

BBC. (2018). Police forces 'struggling' to grasp social media [online]. Last accessed 12th August 2015 at: https://www.bbc.co.uk/news/uk-wales-41127674
BBC. (2016a). *Thousands protest against austerity*. [online]. Last accessed 3rd of May 2019 at: https://www.bbc.co.uk/news/uk-36063743

BBC. (2016b). Far-right and anti-racism protesters clash in Dover [online]. Last accessed 3rd of May 2019 at: https://www.bbc.co.uk/news/uk-england-kent-35450115

BBC. (2016c). *Million Mask March sees 53 arrests*. [online]. Last accessed 3rd of May 2019 at: https://www.bbc.co.uk/news/uk-england-london-37886876

BBC. (2015). *Police hurt but march 'mainly peaceful'*. [online] Last accessed 3rd of May 2019 at: https://www.bbc.co.uk/news/uk-england-london-34743083

Bermingham A., et al. (2009). *Combining social network analysis and sentiment analysis to explore the potential for online radicalization*. ASONAM

Blomberg, J. (2012). Twitter and Facebook Analysis: It's Not Just for Marketing Anymore [online]. Last accessed 12th August 2015 at: http://support.sas.com/resources/papers/proceedings12/309-2012.pdf

Birchley, E. (2015). Big Increase In Facebook And Twitter Crimes [online]. Last accessed 12th August 2015 at: http://news.sky.com/story/1496511/big-increase-in-facebook-and-twitter-crimes
Bobicev, V., and Sokolova, M. (2017). Inter-Annotator Agreement in Sentiment Analysis: Machine Learning Perspective. *RANLP*.

Boom, C.D., Steven V.C., Bart, D. (2015). Semantics-driven Event Clustering in Twitter Feeds [online]. Last accessed 12th August 2015 at: http://ceur-ws.org/Vol-1395/microposts2015_proceedings.pdf#page=41

Borgatti, S., Everett, M., and Johnson, J. (2018). *Analyzing social networks*. 2nd edition. SAGE Publications Ltd.

Borges, J. (2011). Knowledge Extraction and Machine Learning [online]. Last accessed 12th August 2015 at: http://paginas.fe.up.pt/~ec/files_1112/week_01_introductiontoDM.pdf
Borges, J. (2004). CRISP-DM [online]. Last accessed 12th August 2015 at: http://paginas.fe.up.pt/~ec/files_0405/slides/02%20CRISP.pdf

Boyd, D., and Crawford, K. (2012). CRITICAL QUESTIONS FOR BIG DATA: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society,* **15** (5), 662-679.

Brain, T. (2010). A History of Policing from 1974: The Turbulent Years. First Edition. Oxford University Press.

Brown, M. (2014). *Data Mining For Dummies*. 1st Edition. For Dummies.

Bryman, A. (2012). *Social Research Methods*. 4th Edition. Oxford University Press, USA.

Bullock, S. (2015). Reshaping policing for the public [online]. Last accessed 24th August 2015 at: https://policinginsight.com/analysis/reshaping-policing-for-the-public/

Bullock, K. (2018). The Police Use of Social Media: Transformation or Normalisation? *Social Policy and Society, 17*(2), 245-258. Last accessed 24th August 2015 at: doi:10.1017/S1474746417000112

Burnap, P., et al. (2015). COSMOS: Towards an integrated and scalable service for analysing social media on demand. *International journal of parallel, emergent and distributed systems,* **30** (2), 80-100.

Bruns, A. (2017). Challenges in Social Media Research Ethics [online]. Last accessed 24th August 2017 at: http://snurb.info/node/2227

Cambria, E. (2013). An introduction to concept-level sentiment analysis. In: *Advances in Soft Computing and Its Applications.* Springer, 478-483.

Cano A.E., et al. (2013). A weakly supervised bayesian model for violence detection in social media. In: *Proceeding of IJCNLP*

Centre for Data Ethics and Innovation (CDEI). (2020a). Review into bias in algorithmic decision-making [online]. Last accessed 5th August 2021: https://www.gov.uk/government/publications/cdei-publishes-review-into-bias-in-algorithmic-decision-making/main-report-cdei-review-into-bias-in-algorithmic-decision-making

Centre for Data Ethics and Innovation (CDEI). (2020b). Snapshot Series Facial Recognition Technology [online]. Last accessed 5th August 2021 at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/905267/Facial_Recognition_Technology_Snapshot_UPDATED.pdf

Chakraborty, G., Pagolu, M., and Garla, S. (2013). Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS. First Edition. SAS Institute.

Challenger, R., et al. (2009). Understanding Crowd Behaviours: Supporting Evidence [online]. Last accessed 24th August 2015 at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/192606/understanding_crowd_behaviour-supporting-evidence.pdf

Chan, E., and Dobuzinskis, A. (2014). U.S. police struggle to uncover threats on social media [online]. Last accessed 12th August 2015 at: http://www.reuters.com/article/2014/12/26/us-usa-police-socialmedia-idUSKBN0K40MD20141226

Choudhury, M.D., et al. (2016). Social Media Participation in an Activist Movement for Racial Equality. *Proceedings of the International AAAI Conference on Weblogs and Social Media. International AAAI Conference on Weblogs and Social Media, 2016*, 92-101.

Christie, L. (2021). AI in policing and security [online]. Last accessed 5th August 2021 at: https://post.parliament.uk/ai-in-policing-and-security/

Cohen K., et al. (2014) Detecting linguistic markers for radical violence in social media. *Terror Polit Violence* 26(1):246–256

Cohen, L., Manion, L., and Morrison, K. (2011). *Research Methods in Education*. 7th Edition. Routledge.

College of Policing. (2018). Public order: Planning and deployment [online]. Last accessed 12th August 2015 at: https://www.app.college.police.uk/app-content/public-order/planning-and-deployment/?s=disorder+model

College of Policing. (2015). POLKA [online]. Last accessed 12th August 2015 at: http://www.college.police.uk/What-we-do/Research/polka/Pages/POLKA.aspx

College of Policing. (2014). Command [online]. Last accessed 15th May 2015 at: https://www.app.college.police.uk/app-content/public-order/command/

College of Policing. (2013a). Public order: Communication [online]. Last accessed 5th March 2017 at: https://www.app.college.police.uk/app-content/public-order/planning-and-deployment/communication/

College of Policing. (2013b). Command structures [online]. Last accessed 15th May 2015 at: https://www.app.college.police.uk/app-content/operations/command-and-control/command-structures/

Consortiumnews. (2014). NSA Insiders Reveal What Went Wrong [online]. Last accessed 12th August 2015 at: https://consortiumnews.com/tag/edward-loomis?print=print-page

CNN. (2017). Facebook tops 1.9 billion monthly users [online]. Last accessed 12th August 2017 at: http://money.cnn.com/2017/05/03/technology/facebook-earnings/index.html

Constable, N. (2015a). Disconnected: The future of the Police Federation [online]. Last accessed 24th August 2015 at: https://policinginsight.com/opinion/disconnected-the-future-of-the-police-federation/

Constable, N. (2015b). Peel versus The Public? [online]. Last accessed 24th August 2015 at: https://policinginsight.com/opinion/peel-versus-the-public/

Constable, N. (2015c). At all costs [online]. Last accessed 24th August 2015 at: https://policinginsight.com/opinion/at-all-costs/

Cook, D., et al. (2014). Twitter Deception and Influence: Issues of Identity, Slacktivism, and Puppetry. *Journal of Information Warfare, 13*(1), 58-71. Last accessed 24th August 2015 at: https://www.jstor.org/stable/26487011

Couchman, A. (2018). Dover is peaceful now' say residents as a two-year investigation into the infamous riots finally concludes [online]. Last accessed 24th August 2018 at: https://www.kentlive.news/news/kent-news/dover-peaceful-now-say-residents-1667108

Cowley, R. (2006). The history of Her Majesty's Inspectorate of Constabulary [online]. Last accessed 24th August 2015 at: https://www.justiceinspectorates.gov.uk/hmic/media/the-history-of-hmic-the-first-150-years.pdf

Crown Prosecution Service (CPS). (2015). Public Order Offences incorporating the Charging Standard [online]. Last accessed 15th May 2015 at: http://www.cps.gov.uk/legal/p_to_r/public_order_offences/

DataCamp. (2014). Statistical Language Wars [online]. Last accessed 12th August 2015 at: http://datacamp.wpengine.com/wp-content/uploads/2014/05/infograph.png

Dearden, L. (2017). How technology is allowing police to predict where and when crime will happen [online]. Last accessed 12th August 2015 at: https://www.independent.co.uk/news/uk/home-news/police-big-data-technology-predict-crime-hotspot-mapping-rusi-report-research-minority-report-a7963706.html

DeCastella, T., and McLatchy, C. (2011). UK riots: What turns people into looters? [online]. Last accessed 12th August 2015 at: http://www.bbc.co.uk/news/magazine-14463452

Della Porta, D. (1995). Social Movements, Political Violence, and the State: A Comparative Analysis of Italy and Germany. Cambridge: *Cambridge University Press*

Denef, S. (2013). Social media and the police: Tweeting practices of British police forces during the august 2011 riots. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* ACM, 3471-3480.

Dencik, L., Hintz, A., and Carey, Z. (2018). Prediction, pre-emption and limits to dissent: Social media and big data uses for policing protests in the United Kingdom. *New Media & Society*, 20(4), 1433–1450. Last accessed 10th December 2018 at: doi: https://doi.org/10.1177/1461444817697722

Dencik, L., et al. (2015). *Managing 'Threats': Uses of Social Media for Policing Domestic Extremism and Disorder in the UK*. [online] Orca.cf.ac.uk. Last accessed 10th December 2018 at: http://orca.cf.ac.uk/85618/1/Managing-Threats-Project-Report.pdf

de Vries, A. (2012). *R For Dummies*. First Edition. For Dummies.

Domo. (2015). Modern BI For All [online]. Last accessed 10th December 2018 at: https://www.domo.com/

D'Onfro, J. (2013). Twitter Admits 5% Of Its 'Users' Are Fake [online]. Last accessed 10th December 2018 at: http://www.businessinsider.com/5-of-twitter-monthly-active-users-are-fake-2013-10

Droba, D.D. (1931). Methods Used for Measuring Public Opinion. *American Journal of Sociology,* 37 (1931): 410-423.

Drury, J., and Reicher, S. (2000) Collective action and psychological change: the emergence of new social identities. *British Journal of Social Psychology*.

Dul, J., and Hak, T. (2007). *Case Study Methodology in Business Research*. Butterworth-Heinemann, Oxford.

Durham Police. (2017). Peels Principles of Law Enforcement [online]. Last accessed 15th August 2018 at: https://www.durham.police.uk/About-Us/Documents/Peels_Principles_Of_Law_Enforcement.pdf

Eadson, W. (2011). Community Mapping and Tension Monitoring: A practical guide and sourcebook of information and ideas for Welsh local authorities and their partners [online]. Last accessed 15th May 2015 at: http://www.shu.ac.uk/research/cresr/sites/shu.ac.uk/files/community-mapping-tension-monitoring.pdf

EMC. (2015). Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. 1st Edition. *John Wiley & Sons*.

Endsley, M.R. (1988). Situation awareness global assessment technique (SAGAT). In: *Aerospace and electronics conference, 1988. NAECON 1988., proceedings of the IEEE 1988 national,* 789-795 vol.3.

ESRC. (2015). Framework for research ethics [online]. Last accessed 15th August 2015 at: http://www.esrc.ac.uk/_images/framework-for-research-ethics_tcm8-33470.pdf

Essers, L. (2013). U.K. Police: six citizens were wrongly detained due to data mining errors last year [online]. Last accessed 12th August 2015 at: http://www.pcworld.com/article/2044747/bad-internet-data-requests-led-to-6-wrongly-held-or-accused-in-uk.html

Esuli, A., and Sebastiani, F. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. *LREC*.

Evans, H., Ginnis, S., and Bartlett, J. (2015). #SocialEthics a guide to embedding ethics in social media research [online]. Last accessed 10th December 2015 at:

https://www.ipsos.com/sites/default/files/migrations/en-uk/files/Assets/Docs/Publications/im-demos-social-ethics-in-social-media-research-summary.pdf

Fan, W., and Bifet, A. (2013). Mining big data: current status, and forecast to the future. *ACM SIGKDD explorations newsletter,* **14** (2), 1.

Feinerer, I. (2015). Package 'tm' [online]. Last accessed 24th August 2015 at: https://cran.r-project.org/web/packages/tm/tm.pdf

Fernandez, M., Dickinson, T., and Alani, H. (2017). An analysis of UK Policing Engagement via Social Media. In: Social Informatics. SocInfo 2017 (Ciampaglia, G.; Mashhadi, A. and Yasseri, T. eds.), Lecture Notes in Computer Science. *Springer*, pp. 289–304.

Fichet, E.S., et al. (2016). Eyes on the Ground: Emerging Practices in Periscope Use during Crisis Events. *ISCRAM*.

Flores-Ruiz, D., Elizondo-Salto, A., & Barroso-González, M.D. (2021). Using Social Media in Tourist Sentiment Analysis: A Case Study of Andalusia during the Covid-19 Pandemic. *Sustainability, 13*, 3836.

García, A., Gaines, S., and Linaza, M. (2012). A Lexicon based sentiment analysis retrieval system for tourism domain. *ICIT 2012*.

Gartner. (2015). Magic Quadrant for Advanced Analytics Platforms [online]. Last accessed 12th August 2015 at: http://www.gartner.com/technology/reprints.do?id=1-2A881DN&ct=150219&st=sb

Gate. (2019). Performance Evaluation of Language Analysers [online]. Last accessed 28th March 2019 at: https://gate.ac.uk/sale/tao/splitch10.html

Gayle, D. (2016a). Million Mask March: police curb protests amid fears of violence [online]. Last accessed 3rd May 2019 at: https://www.theguardian.com/technology/2016/nov/04/million-mask-march-police-protesters-violence-fears-restrictions

Gayle, D. (2016b). Far-right and anti-fascist protesters clash in Dover [online]. Last accessed 3rd May 2019 at: https://www.theguardian.com/uk-news/2016/jan/30/far-right-anti-fascist-protesters-clash-dover

Gayle, D. and Johnston, C. (2015a). Million Mask march: scores arrested after clashes between police and protesters [online]. Last accessed 3rd May 2019 at:

https://www.theguardian.com/uk-news/2015/nov/05/three-arrested-anti-capitalist-protesters-million-mask-march

Gayle, D., and Johnston, C. (2015b). Million Mask march in London - as it happened [online]. Last accessed 3rd May 2019 at: https://www.theguardian.com/uk-news/live/2015/nov/05/million-mask-march-gathers-in-london-live-updates

Georgakopoulou, A., Iversen, S., and Stage, C. (2020). *Quantified Storytelling: A Narrative Analysis of Metrics on Social Media*. First edition. Palgrave Macmillan.

Gerbaudo, P. (2017). Social media teams as digital vanguards: the question of leadership in the management of key Facebook and Twitter accounts of Occupy Wall Street, Indignados and UK Uncut. *Information, Communication & Society*. 20 (2), 185–202.

Gov. (2021). HM Inspectorate of Constabulary [online]. Last accessed 5th August 2021: https://www.gov.uk/government/organisations/hm-inspectorate-of-constabulary

Glass, K., and Colbaugh, R. (2011). Web Analytics for Security Informatics. *2011 European Intelligence and Security Informatics Conference*, 214-219.

Gorringe, H., Stott, C., and Rosie, M. (2012). Dialogue Police, Decision Making, and the Management of Public Order During Protest Crowd Events. *Journal of investigative psychology and offender profiling,* **9** (2), 111-125.

Government Social Research (GSR). (2016). Using social media for social research: An introduction [online]. Last accessed 3rd March 2018 at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/524750/GSR_Social_Media_Research_Guidance_-_Using_social_media_for_social_research.pdf

GlobalWebIndex (GWI). (2014). GWI Social Global WebIndex's Quarterly Report on The Latest Trends in Social Networking Q4 2014 [online]. Last accessed 12th August 2015 at: http://www.slideshare.net/globalwebindex/gwi-social-report-q4-2014

GlobalWebIndex (GWI). (2015). UK'S Top Social Networks GlobalWebIndex Infographics, Q1 2015 [online]. Last accessed 12th August 2015 at: http://pro.globalwebindex.net/#/infographics/17937

Global Web Index (GWI). (2016). *SOCIAL SUMMARY: GlobalWebIndex's quarterly report on the latest trends in social networking* [online]. Last accessed 27th November 2018 at: http://insight.globalwebindex.net/hubfs/Reports/GWI-Social-Q4-2016-Summary-Report.pdf

Greer, C., and Mclauglin, E. (2010). We Predict a Riot?: Public Order Policing, New Media Environments and the Rise of the Citizen Journalist. *British journal of criminology,* **50** (6), 1041-1059.

Grierson, J. (2016). Anti-austerity protest: tens of thousands attend London march [online]. Last accessed 3rd May 2019 at: https://www.theguardian.com/world/2016/apr/16/london-anti-austerity-march-draws-tens-of-thousands

Gutteridge, N., and Wood, V. (2016). 'Death to the monarchy' Angry scenes as anarchists invade London for Million Mask March [online]. Last accessed 3rd May 2019 at: https://www.express.co.uk/news/uk/729179/Million-Mask-March-2016-Anonymous-anarchists-protest-London-UK

Hansard, D. (2013). Written Answers to Questions [online]. Last accessed 12th August 2015 at: http://www.publications.parliament.uk/pa/cm201314/cmhansrd/cm131209/text/131209w0001.htm

Harbisher, B. (2016). The Million Mask March: Language, legitimacy, and dissent. *Critical Discourse Studies*. 13 (3), 1–16.

Harvard. (2002). General Inquirer Categories [online] Last accessed 28th March 2019 at: http://www.wjh.harvard.edu/~inquirer/homecat.htm

Her Majesty (HM) Government. (2011). Contest The United Kingdom's Strategy for Countering Terrorism [online]. Last accessed 2nd December 2015 at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/97995/strategy-contest.pdf

Her Majesty's Inspectorate of Constabulary (HMIC). (2015). Policing in Austerity: Meeting the Challenge [online]. Last accessed 12th August 2015 at: http://www.justiceinspectorates.gov.uk/hmic/wp-content/uploads/policing-in-austerity-meeting-the-challenge.pdf

Her Majesty's Inspectorate of Constabulary HMIC. (2012). Revisiting Police Relationships [online]. Last accessed 12th August 2015 at: https://www.justiceinspectorates.gov.uk/hmic/media/revisiting-police-relationships.pdf

Her Majesty's Inspectorate of Constabulary (HMIC). (2011a). Policing public order: An overview and review of progress against the recommendations of adapting to protest and nurturing the British model of policing [online]. Last accessed 15th May 2015 at:

http://www.justiceinspectorates.gov.uk/hmic/media/policing-public-order-20110208.pdf

Her Majesty's Inspectorate of Constabulary (HMIC). (2011b). The rules of engagement: A review of the August 2011 disorders [online]. Last accessed 15th May 2015 at: http://www.justiceinspectorates.gov.uk/hmic/media/a-review-of-the-august-2011-disorders-20111220.pdf

Her Majesty's Inspectorate of Constabulary (HMIC). (2009). Adapting to Protest [online]. Last accessed 12th August 2015 at: http://www.justiceinspectorates.gov.uk/hmic/media/adapting-to-protest-20090705.pdf

Her Majesty's Inspectorate of Constabulary (HMIC). (2006). The history of Her Majesty's Inspectorate of Constabulary [online]. Last accessed 12th August 2015 at: https://www.justiceinspectorates.gov.uk/hmic/media/the-history-of-hmic-the-first-150-years.pdf

Her Majesty's Inspectorate of Constabulary and Fire & Rescue Services (HMICFRS). (2021). Strategy 2021–25 [online]. Last accessed 5th August 2021: https://www.justiceinspectorates.gov.uk/hmicfrs/wp-content/uploads/hmicfrs-strategy-2021-25.pdf

Hollywood, J., et al. (2018). Using Social Media and Social Network Analysis in Law Enforcement: Creating a Research Agenda, Including Business Cases, Protections, and Technology Needs.

Homeland Security. (2014). Using Social Media for Enhanced Situational Awareness and Decision Support [online]. Last accessed 15th May 2015 at: http://www.firstresponder.gov/TechnologyDocuments/Using%20Social%20Media%20for%20Enhanced%20Situational%20Awareness%20and%20Decision%20Support.pdf

Home Office. (2015). Have you got what it takes? Dealing with public order [online]. Last accessed 15th May 2015 at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/117436/dealing-with-public-order.pdf

Home Office. (2012). Policing by consent [online]. Last accessed 9th November 2015 at: https://www.gov.uk/government/publications/policing-by-consent

House of Commons. (2018). Policing for the future [online]. Last accessed 5th August 2021 at: https://publications.parliament.uk/pa/cm201719/cmselect/cmhaff/515/515.pdf

House of Commons. (2014a). Serious Organised Crime Agency Annual Report and Accounts [online]. Last accessed 12th August 2015 at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/330516/SOCA2013-14.pdf

House of Commons. (2014b). Social media data and real time analytics [online]. Last accessed 12th August 2015 at: http://www.parliament.uk/business/committees/committees-a-z/commons-select/science-and-technology-committee/news/140612-smd-ev/
http://dl.acm.org/citation.cfm?id=2789198
House of Commons. (2013). 2012 Annual Report of the Interception of Communications Commissioner [online]. Last accessed 12th August 2015 at: http://www.iocco-uk.info/docs/2012%20Annual%20Report%20of%20the%20Interception%20of%20Communications%20Commissioner%20WEB.pdf

House of Commons. (2011). Policing Large Scale Disorder: Lessons from the disturbances of August 2011 [online]. Last accessed 15th May 2015 at: http://www.publications.parliament.uk/pa/cm201012/cmselect/cmhaff/1456/1456i.pdf

Hurwitz, Judith. (2013). *Big Data For Dummies*. 1st Edition. John Wiley & Sons.

IBM. (2015a). IBM SPSS Modeler Text Analytics 15 User's Guide [online]. Last accessed 24th August 2015 at: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/en/Users_Guide_For_Text_Analytics.pdf

IBM. (2015b). SPSS Text Analytics for Surveys [online]. Last accessed 24th August 2015 at: http://www-03.ibm.com/software/products/en/spss-text-analytics-surveys

IBM. (2011). IBM SPSS Modeler CRISP-DM Guide [online]. Last accessed 12th August 2015 at: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf

IBM. (2000). CRISP-DM 1.0 [online]. Last accessed 12th August 2015 at: ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf

Independent Police Commission (IPC). (2013). Policing for a Better Britain Report of the Independent Police Commission [online]. Last accessed 24th August 2015 at:

http://independentpolicecommission.org.uk/uploads/37d80308-be23-9684-054d-e4958bb9d518.pdf

Independent Police Commission (IPC). (2015). Peelian Principles [online]. Last accessed 9th November 2015 at: http://independentpolicecommission.org.uk/peelian-principles

Innes, M., and Roberts, C. (2011). Policing, Situational Intelligence & The Information Environment: A Report To Her Majesty's Inspectorate Of Constabulary [online]. Last accessed 15th May 2015 at: http://issuu.com/upsi/docs/hmic_stint

ITV News. (2016). Anti-austerity march in London attracts '150,000' protesters [online]. Last accessed 3rd May 2019 at: https://www.itv.com/news/2016-04-16/anti-austerity-march-in-london-attracts-150-000-protesters/

Institute of Community Cohesion (iCOCO). (2010) Understanding and monitoring tension and conflict in local communities [online]. Last accessed 15th May 2015 at: http://www.cohesioninstitute.org.uk/live/images/cme_resources/Public/documents/tension%20monitoring%20toolkit%202nd%20edition/tension_monitoring_second_ed.pdf

International Association Of Chiefs Of Police (IACP). (2015). International Association Of Chiefs Of Police 2014 Social Media Survey Results [online]. Last accessed 6th March 2015 at: http://www.iacpsocialmedia.org/Portals/1/documents/2014SurveyResults.pdf

Isaac Newton Institute. (2017). Change-point Detection [online]. Last accessed 6th March 2018 at: https://www.newton.ac.uk/files/seminar/20140113140015001-153901.pdf

Islam, M., and Zibran, M. (2017). A Comparison of Dictionary Building Methods for Sentiment Analysis in Software Engineering Text. *2017 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*.

Janoowalla, M. (2015). Digital policing for a digital age [online]. Last accessed 24th August 2015 at: https://policinginsight.com/analysis/digital-policing-for-a-digital-age-2/

Jensen, M. (2016). Social Media and Political Campaigning. *The International Journal Of Press/Politics*, *22*(1), 23-42. Last accessed 6th March 2018 at: doi: 10.1177/1940161216673196

Jockers, M. (2017). Syuzhet: Extract sentiment and plot arcs from Text [online]. Last accessed 1st April 2019 at: https://github.com/mjockers/syuzhet

Joint Emergency Services Interoperability Programme (JESIP). (2013). Joint Doctrine: the interoperability framework [online]. Last accessed 24th August 2015 at: http://www.jesip.org.uk/wp-content/uploads/2013/07/JESIP-Joint-Doctrine.pdf

Jungherr, A., and Jürgens, P. (2013). Stuttgart's black Thursday on Twitter: Mapping Political Protests with Social Media Data. Last accessed 24th August 2015 at: http://andreasjungherr.net/wp-content/uploads/2008/10/Jungherr-J%C3%BCrgens-Stuttgarts-black-Thursday-on-Twitter-Preprint.pdf

Jurek, A., Mulvenna, M., and Bi, Y. (2015). Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics*, 4(1).

Jurka, T. (2012). Sentiment: Tools for Sentiment Analysis R package [online]. Last accessed 24th August 2015 at: https://cran.r-project.org/package=sentiment

Karmi, O. (2013). David Miranda challenges legality of UK police data mining [online]. Last accessed 12th August 2015 at: http://www.thenational.ae/news/world/europe/david-miranda-challenges-legality-of-uk-police-data-mining

Karpf, D. (2012). *The MoveOn Effect: The Unexpected Transformation of American Political Advocacy* (1st ed.). New York: Oxford University Press.

Karyotis, G., and Rüdig, W. (2018). The Three Waves of Anti-Austerity Protest in Greece, 2010–2015. *Political Studies Review*. 16 (2), 158–169.

Kass-Hout, T., and Xu, Z. (2017). Change Point Analysis [online]. Last accessed 12th August 2018 at: https://sites.google.com/site/changepointanalysis/

Kearns, I., and Muir, R. (2019). Data-Driven Policing and Public Value [online]. Last accessed 5th August 2021 at: https://www.police-foundation.org.uk/2017/wp-content/uploads/2010/10/data_driven_policing_final.pdf

Keith, J., Ginnis, S., and Miller, C. (2016). Addressing quality in social media research: the question of representivity. Last accessed 12th August 2018 at: http://the-sra.org.uk/wp-content/uploads/social-research-practice-journal-issue-02-summer-2016-v2.pdf

Kietzmann, J.H., et al. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business horizons,* **54** (3), 241-251.

Killick, R. (2017). Introduction to optimal changepoint detection algorithms [online]. Last accessed 12th August 2018 at: http://members.cbio.mines-paristech.fr/~thocking/change-tutorial/RK-CptWorkshop.html

Killick, R., and Eckley, I.A. (2014). changepoint: An R package for changepoint analysis. *Journal of Statistical Software* 58(3) 1-19.

King, M., and Waddington, D. (2004). Coping with disorder? the changing relationship between police public order strategy and practice a critical analysis of the Burnley Riot. *Policing and society,* **14** (2), 118-137.

Knott, C., and Steube, G. (2010). Using EXCEL in an introductory statistics course: A comparison of instructor and student [online]. Last accessed 12th August 2015 at: http://www.nedsi.org/proc/2010/proc/p091112001.pdf

Koper, C., Lum, C., and Willis, J. (2014). Optimizing the Use of Technology in Policing: Results and Implications from a Multi-Site Study of the Social, Organizational, and Behavioural Aspects of Implementing Police Technologies. *Policing-an International Journal of Police Strategies & Management, 8*, 212-221.

Korolov, R., et al. (2016). On predicting social unrest using social media. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.

Kotzias, D., et al. (2015). *From group to individual labels using deep features*. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 597-606.

Knight, W. (2018). Anonymous: a social movement [online]. PhD thesis. University of Nottingham. Last accessed 25th August 2019 at: http://eprints.nottingham.ac.uk/46804/1/Full%20Thesis%20-%20William%20Knight%20-%20psxwk2.pdf

Kumar, P. (2013). Law Enforcement and Mining Social Media: Where's the Oversight? [online]. Last accessed 12th August 2015 at: https://blogs.law.harvard.edu/internetmonitor/2013/07/01/law-enforcement-and-mining-social-media-wheres-the-oversight/

Landis, J.R., and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33 1*, 159-74.

Lantz, B. (2015). *Machine learning with R*. 2nd ed. Birmingham: Packt Publ.

Laville, S. (2012). Police forces warned to treat their tweeters with care [online]. Last accessed 12th August 2015 at: http://www.theguardian.com/uk/2012/oct/02/police-warned-treat-tweeters-with-care

Lennon, S. (2016). Dover riots of January 30, 2016 - police hunt for far right and anti-fascist thugs goes on [online]. Last accessed 12th August 2015 at: https://www.kentonline.co.uk/dover/news/dover-riots---one-year-119566/

Lewis, P., et al. (2011). *Reading the riots: investigating England's summer of disorder.* Reading the riots. The London School of Economics and Political Science and The Guardian, London, UK.

LexisNexis. (2014a). Social Media Use in Law Enforcement: Crime prevention and investigative activities continue to drive usage. [online]. Last accessed 6th March 2015 at: http://www.lexisnexis.com/risk/downloads/whitepaper/2014-social-media-use-in-law-enforcement.pdf

LexisNexis. (2014b). Using Social Media to Solve Crimes: A Chief Gives Advice [online]. Last accessed 12th August 2015 at: http://blogs.lexisnexis.com/public-safety/2014/08/using-social-media-to-solve-crimes-a-chief-gives-advice/

LexisNexis. (2012). Law Enforcement Personnel Use of Social Media in Investigations: Summary of Findings [online]. Last accessed 6th March 2015 at: http://www.lexisnexis.com/risk/downloads/whitepaper/Infographic-Social-Media-Use-in-Law-Enforcement.pdf

Lin, T., et al. (2008). *Data Mining: Foundations and Practice*. First edition. Springer Berlin Heidelberg.

Liu, H. (2017). Sentiment analysis of citations using word2vec. *arXiv preprint arXiv:1704.00177*.

Liu, B. (2020). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions (Studies in Natural Language Processing)*. Publisher: Cambridge University Press. 2nd edition.

Liu, B. (2015). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Publisher: *Cambridge University Press*.

Liu, B. (2012). Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technologies). *Morgan & Claypool Publishers*.

Loughran, T., and McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, 54(4), pp.1187-1230.

Mäntylä, M., Graziotin, D., and Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*, *27*, 16-32. Last accessed 6th March 2015 at: doi: 10.1016/j.cosrev.2017.10.002

Mass, P. (2015). Inside NSA, Officials Privately Criticize "COLLECT IT ALL" Surveillance [online]. Last accessed 12th August 2015 at: https://firstlook.org/theintercept/2015/05/28/nsa-officials-privately-criticize-collect-it-all-surveillance/

McCallum, Q. (2013). *Bad data handbook* (1st ed.). Beijing: O'Reilly.
McCarthy, J., and Warner, C. (2014). Whatever can go wrong will: situational complexity and public order policing. [online]. *Policing and society,* **24** (5), 566-587.
McCue, C. (2010). Connecting the Dots: Data Mining and Predictive Analytics in Law Enforcement and Intelligence Analysis.

McCue, C. (2006). Data Mining and Crime Analysis in the Richmond Police Department [online]. Last accessed 12th August 2015 at: http://www.spss.ch/eupload/File/PDF/Data%20Mining%20and%20Crime%20Analysis%20in%20the%20Richmond%20Police%20Departement.pdf

McCue, C. (2003). Connecting the Dots: Data Mining and Predictive Analytics in Law Enforcement and Intelligence Analysis [online]. Last accessed 12th August 2015 at: http://www.policechiefmagazine.org/magazine/index.cfm?fuseaction=display_arch&article_id=121&issue_id=102003

McSeveny, K., and Waddington, D. (2011). Up close and personal: the interplay between information technology and human agency in the policing of the 2011 Sheffield Anti-Lib Dem protes. In: AKHGAR, Babak and YATES, Simeon, (eds.) *Intelligence management : knowledge driven frameworks for combating terrorism and organized crime.* Advanced Information and Knowledge Processing. Springer, 199-212.

McPhail, C., Schweingruber, D., and McCarthy, J. (1998). 'Policing Protest in the United States: 1960–1995', in D. della Porta and H. Reiter (eds). Policing Protest: The Control of Mass Demonstrations in Western Democracies. Minneapolis: *University of Minnesota Press*, pp. 49–69.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093-1113. Last accessed 12th August 2018 at: doi: 10.1016/j.asej.2014.04.011

Metropolitan Police Service (MPS). (2021). Met Strategic Digital Enabling Framework 2021-25 [online]. Last accessed 5th August 2021 at: https://www.met.police.uk/SysSiteAssets/media/downloads/force-content/met/about-us/met-strategic-digital-enabling-framework-2021-2025.pdf

Metropolitan Police Service (MPS). (2014). Digital Policing: Technology that enables crime fighting, improves victim care and reduces the cost of operations [online]. Last accessed 10th November 2014 at: http://content.met.police.uk/cs/Satellite?blobcol=urldata&blobheadername1=Content-Type&blobheadername2=Content-Disposition&blobheadervalue1=application%2Fpdf&blobheadervalue2=inline%3B+filename%3D%22140%2F125%2FTotal+Technology+Strategy+-+2014-2017.pdf%22&blobkey=id&blobtable=MungoBlobs&blobwhere=1283686449257&ssbinary=true

Metropolitan Police Service (MPS). (2012). 4 Days in August Strategic Review into the Disorder of August 2011 [online]. Last accessed 10th November 2014 at: http://www.met.police.uk/foi/pdfs/priorities_and_how_we_are_doing/corporate/4_days_in_august.pdf

Metropolitan Police (MET). (2015). MPS historical timeline [online]. Last accessed 12th August 2015 at: http://content.met.police.uk/Site/historicaltimeline

Milan, S., & van der Velden, L. (2018). Data Activism. Krisis : Journal for contemporary philosophy, 2018(1). http://krisis.eu/issue-1-2018-data-activism/

Miller, C., et al. (2015). the road to representivity a Demos and Ipsos MORI report on sociological research using Twitter [online]. Last accessed 12th August 2015 at: http://www.demos.co.uk/files/Road_to_representivity_final.pdf?1441811336

Mittal, A., and Goel, A. (2013). Stock prediction using twitter sentiment analysis. In: *Proceeding of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*

Mohammad, S., and Turney, P. (2010). Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon. *In Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text (CAAGET '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 26-34.

Moreno, M.A., et al. (2013). Ethics of social media research: common concerns and practical considerations. *Cyberpsychology, behavior and social networking,* **16** (9), 708-713.

Morozov, E. (2013). How Facebook Could Get You Arrested [online]. Last accessed 12th August 2015 at: https://zcomm.org/znetarticle/how-facebook-could-get-you-arrested-by-evgeny-morozov/

Mudinas, A., Zhang, D., and Levene, M. (2012). Combining lexicon and learning based approaches for concept-level sentiment analysis. *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining.* ACM. New York, NY, USA, Article 5, pp. 1-8.

Murji, K., and Neal, S. (2011). Riot: Race and Politics in the 2011 Disorders. *Sociological Research Online*, *16*(4), 216–220. Last accessed 12th August 2015 at: doi: https://doi.org/10.5153/sro.2557

NatCen. (2014). Research using Social Media; Users' Views [online]. Last accessed 12th August 2015 at: http://www.natcen.ac.uk/media/282288/p0639-research-using-social-media-report-final-190214.pdf

Nagesh, A. (2016a). Guy Fawkes masked protesters in London for Million Mask March | Metro News [online]. Last accessed 3rd May 2019 at: https://metro.co.uk/2016/11/05/thousands-of-protesters-in-guy-fawkes-masks-take-over-london-for-million-mask-march-6237950/

Nagesh, A. (2016b). 'Neo-Nazi gangs paint blood swastikas' at violent clash with anti-fascists in Dover [online]. Last accessed 3rd May 2019 at: https://metro.co.uk/2016/01/30/huge-neo-nazi-gangs-paint-blood-swastikas-at-massive-demonstration-in-dover-5653205/

National Crime Agency (NCA). (2015). About us [online]. Last accessed 6th March 2015 at: http://www.nationalcrimeagency.gov.uk/about-us

National Crime Agency (NCA). (2013). Suspicious Activity Reports (SARs) Annual Report 2013 [online]. Last accessed 12th August 2015 at: http://www.nationalcrimeagency.gov.uk/publications/94-sars-annual-report-2013/file

National Police Chiefs' Council (NPCC). (2015a). Chief Constable Sara Thornton has been appointed as Chair of the National Police Chiefs' Council (NPCC) [online]. Last accessed 12th August 2015 at: http://news.npcc.police.uk/releases/chief-constable-sara-thornton-has-been-appointed-as-chair-of-the-national-police-chiefs-council-npcc

National Police Chiefs' Council (NPCC). (2015b). The National Community Tension Team [online]. Last accessed 12th August 2015 at: http://www.npcc.police.uk/NPCCBusinessAreas/PREVENT/TheNationalCommunityTensionTeam.aspx

National Police Chiefs' Council (NPCC). (2015c). History and background [online]. Last accessed 24th August 2015 at: http://www.npcc.police.uk/About/History.aspx

National Police Chiefs' Council (NPCC). (2014). Social media is a powerful tool for police to engage with the public [online]. Last accessed 24th August 2015 at: http://news.npcc.police.uk/releases/social-media-is-a-powerful-tool-for-police-to-engage-with-the-public

Neuropolitics. (2016). UK-EU Twitter Sentiment Analysis: An analysis of the sentiment in the twittersphere towards the UK leaving or remaining in the EU [online]. Last accessed 24th August 2015 at: https://blogs.sps.ed.ac.uk/neuropolitics/2016/01/06/uk-eu-twitter-sentiment-analysis-an-analysis-of-the-sentiment-in-the-twittersphere-towards-the-uk-leaving-or-remaining-in-the-eu/

Nickerson, J. (2016). Thousands turn out in Trafalgar Square for anti-austerity march [online]. Last accessed 12th August 2021 at: https://www.cityam.com/thousands-turn-out-in-trafalgar-square-for-anti-austerity-march/

Nielsen, F.Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

N8 Policing Research Partnership. (2015). N8 Policing Research Partnership: Innovation Forum on Cybercrime - Market Place Discussions [online]. Last accessed 12th August 2015 at: http://n8prp.org.uk/wp-content/uploads/2015/11/SocialMedia_Notes.pdf

O'Connor, D. (2010). Her Majesty's Inspectorate of Constabulary in 2009/10 [online]. Last accessed 12th August 2015 at: http://www.justiceinspectorates.gov.uk/hmic/media/annual-report-2009-10.pdf

Olson, D., and Delen, L. (2008). Advanced Data Mining Techniques. Berlin, Heidelberg: *Springer Berlin Heidelberg*.

Omand, D., Bartlett, J., & Miller, C. (2012). Introducing Social Media Intelligence (SOCMINT). *Intelligence and National Security, 27*, 801 - 823.

Office for National Statistics (ONS). (2021). Home internet and social media usage [online]. Last accessed 12th August 2021 at: https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage

Office for National Statistics (ONS). (2016). Social Media in the UK [online]. Last accessed 12th August 2018 at: https://www.slideshare.net/statisticsONS/social-media-use-in-the-uk-65266896

Open Source Communications Analytics Research (OSCAR). (2018). Open Source Communications Analytics Research [OSCAR] Development Centre: FINAL REPORT [online]. Last accessed 12th December 2018 at: https://www.cardiff.ac.uk/__data/assets/pdf_file/0003/921315/OSCAR-Final-Report-Exec-Summary.pdf

Phillips, A. (2016). Social media is changing the face of politics – and it's not good news [online]. Last accessed 12th December 2018 at: https://theconversation.com/social-media-is-changing-the-face-of-politics-and-its-not-good-news-54266

Pickles, C. (2015). The current model of policing? It's no longer fit for purpose [online]. Last accessed 24th August 2015 at: https://policinginsight.com/opinion/the-current-model-of-policing-its-no-longer-fit-for-purpose/

Preotiuc-Pietro, D. (2014). Temporal models of streaming social media data [online]. Last accessed 24th August 2015 at: http://etheses.whiterose.ac.uk/6379/1/PhD_Thesis_Preotiuc.pdf

Procter, R., Vis, F., and Voss, A. (2013). Reading the riots on Twitter: methodological innovation for the analysis of big data. *International Journal of Social Research Methodology, 16*, 197 - 214.

Pruim, R. (2012). Computational Statistics using R and RStudio [online]. Last accessed 12th August 2015 at: http://www.calvin.edu/~rpruim/talks/Rminis/RAdvantages.pdf

Public Opinion Quarterly. (2018). About the Journal [online]. Last accessed 24th August 2018 at: https://academic.oup.com/poq/pages/About

Rees, A. (2015). *Digital and Online Activism | Responsibility* [online]. Last accessed 2nd December 2018 at: https://en.reset.org/knowledge/digital-and-online-activism

Rinker, T. (2017). *stansent* [online]. Last accessed 1st April 2019 at: https://github.com/trinker/stansent

Rkaina, S. (2016). Dover rally: Live updates as far-right groups clash with anti-fascist protesters [online]. Last accessed 15th April 2019 at: https://www.mirror.co.uk/news/uk-news/dover-rally-live-updates-far-7275371

Roberts, D. (2011). Improving Situational Awareness [online]. Last accessed 15th May 2015 at: http://www.policechiefmagazine.org/magazine/index.cfm?fuseaction=print_display&article_id=2485&issue_id=92011

Rogers, S., Sedghi, A., and Evans, L. (2011). UK riots: every verified incident - interactive map [online]. Last accessed 15th April 2019 at: http://www.theguardian.com/news/datablog/interactive/2011/aug/09/uk-riots-incident-map

Saif, H., et al. (2014). On stopwords, filtering and data sparsity for sentiment analysis of Twitter. In: *LREC 2014, Ninth International Conference on Language Resources and Evaluation. Proceedings.*, pp. 810–817.

Sakaki, T., Toriumi, F., and Matsuo, Y. (2011). Tweet trend analysis in an emergency situation. *SWID '11*.

Santos, J., Bernardini, F., and Paes, A. (2021). Measuring the Degree of Divergence when Labeling Tweets in the Electoral Scenario. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, (pp. 127-138). Porto Alegre: SBC.

SAS. (2015a). Using technology to keep the community safe [online]. Last accessed 12th August 2015 at: http://www.sas.com/en_us/customers/west-midlands-police.html

SAS. (2015b). History [online]. Last accessed 12th August 2015 at: http://www.sas.com/en_us/company-information.html#history

SAS. (2015c). SAS Products and Solutions [online]. Last accessed 12th August 2015 at: http://support.sas.com/software/index.html

Scavetta, R., and Angelov, B. (2021). *Python and R for the Modern Data Scientist: The Best of Both Worlds*. Publisher: O'Reilly Media, Inc, USA.

Scherman, A., Arriagada, A., and Valenzuela, S. (2015). Student and Environmental Protests in Chile: The Role of Social Media. *Politics; politics,* **35** (2), 151-171.

Scott, S. (2012). Gagged! Cops banned from tweets on the beat [online]. Last accessed 24th August 2015 at: http://www.northampton-news-hp.co.uk/Gagged-Cops-banned-tweets-beat/story-21672379-detail/story.html

Seawright, J., and Gerring, J. (2008). *Case selection techniques in case study research: a menu of qualitative and quantitative options*. Political Research Quarterly, Vol. 61 No. 2, pp. 294-308.

Selivanov, D., Bickel, M. & Wang, Q., 2020. *Package 'text2vec' [online]*. Last accessed 12th November 2021 at: https://cran.r-project.org/web/packages/text2vec/text2vec.pdf

Selivanov, D., Bickel, M., & Wang, Q. (2020). *Package 'text2vec' [online]*. Last accessed 12th November 2021 at: https://cran.r-project.org/web/packages/text2vec/text2vec.pdf

Silge, J., and Robinson, D. (2017). *Text mining with R* (1st ed.). Beijing: O'Reilly.

Sims, A. (2016). Thousands join Million Mask March in central London [online]. Last accessed 3rd May 2019 at: https://www.independent.co.uk/news/uk/home-news/thousands-join-million-mask-march-in-central-london-after-police-impose-strict-restrictions-a7400021.html

SkyNews. (2015). Big Increase In Facebook And Twitter Crimes [online]. Last accessed 12th August 2015 at: https://twitter.com/SkyNewsTonight/status/606895700371046400

Sloan, L., and Quan-Haase, A (eds.). (2016). *The Sage handbook of social media research methods* (First edition.). SAGE Inc.

Snelson, C. (2016). Qualitative and Mixed Methods Social Media Research. *International Journal of Qualitative Methods*, 15(1), p.160940691562457.

socialdatalab. (2017). Lab Online Guide to Social Media Research Ethics [online]. Last accessed 12th August 2018 at: http://socialdatalab.net/ethics-resources

Social Media Data Stewardship (SMDS). (2017). Social Media Data Stewardship – Homepage [online]. Last accessed 12th August 2018 at: http://socialmediadata.org/

Social Media Ltd. (2016). Most Popular Social Networks in the UK [online]. Last accessed 12th August 2018 at: https://social-media.co.uk/list-popular-social-networking-websites

Solarte, J. (2002). A Proposed Data Mining Methodology and its Application to Industrial Engineering [online]. Last accessed 6th March 2015 at: http://trace.tennessee.edu/cgi/viewcontent.cgi?article=3549&context=utk_gradthes

Statista. (2017). Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2019 [online]. Last accessed 6th August 2018 at: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Stott, C., and Drury, J. (2000). Crowds, context and identity: Dynamic categorization processes in the 'poll tax riot'. *Human Relations, 53*, 247 - 273.

Stott, C., Scothern, M., and Gorringe, H. (2013). Advances in Liaison Based Public Order Policing in England: Human Rights and Negotiating the Management of Protest? *Policing: A journal of policy and practice,* **7** (2), 210-224.

Stupnikov, A. (2014). Introduction on to R: basics [online]. Last accessed 12th August 2015 at: http://www.bio-complexity.com/QUBsscb14/media/Lecture2.pdf

Syracuse University. (2014). Dave Karpf visits TNGO initative [online]. Last accessed 12th August 2015 at: https://www.youtube.com/watch?v=9ZnArPrxCEw

Technical University of Denmark. (2012). Introduction to (parts of) SAS [online]. Last accessed 12th August 2015 at: http://staff.pubhealth.ku.dk/~lts/engelsk_basal/overheads/sas_intro.pdf

The Guardian. (2016). Million Mask March ends with dozens arrested in central London [online]. Last accessed 3rd May 2019 at: https://www.theguardian.com/technology/2016/nov/06/dozens-of-arrests-at-million-mask-march-in-central-london

Thelwall, M. (2019). SentiStrength - sentiment strength detection in short texts - sentiment analysis, opinion mining [online]. Last accessed 1st April 2019 at: http://sentistrength.wlv.ac.uk/

Theocharis, Y., et al. (2014) Using Twitter to mobilize protest action: online mobilization patterns and action repertoires in the Occupy Wall Street, Indignados, and Aganaktismenoi movements. *Information, Communication & Society*. 18 (2), 1–19.

The Open University. (2009). Police and Public Order [online]. Last accessed 15th May 2015 at: http://www.open.ac.uk/Arts/history-from-police-archives/Met6Kt/PublicOrder/poIntro.html

The R Foundation. (2015). What is R? [online]. Last accessed 12th August 2015 at: https://www.r-project.org/about.html

Torre, P. (2018). MEDI@4SEC Workshop 4 Report: Everyday Security Deliverable 2.6 [online]. Last accessed 25th August 2018 at: http://media4sec.eu/downloads/d2-6.pdf

Trottier, D. (2015). Coming to terms with social media monitoring: Uptake and early assessment. *Crime, media, culture,* 1741659015593390.

Turner, C., and Finnigan, L. (2015). Million Mask March: Three officers and six police horses hurt on night of violence in London [online]. Last accessed 3rd May 2019 at:

https://www.telegraph.co.uk/news/uknews/crime/11975183/Million-Mask-March-Anonymous-protesters-hurl-fireworks-at-police-in-London-live.html

Twitter. (2017a). Developer Agreement and Policy [online]. Last accessed 3rd May 2017 at: https://dev.twitter.com/overview/terms/agreement-and-policy

Twitter. (2017b). Twitter Usage / Company Facts [online]. Last accessed 3rd May 2017 at: https://about.twitter.com/company

Unite Community. (2016). March for Health, Homes, Jobs and Education – End Austerity Now! People's Assembly Against Austerity Saturday 16th April London [online]. Last accessed 3rd May 2017 at: https://unitecommunityleedswakefield.wordpress.com/2016/04/07/march-for-health-homes-jobs-and-education-end-austerity-now-peoples-assembly-against-austerity-saturday-16th-april-london/

United States Coast Guard. (2014). Situational Awareness [online]. Last accessed 15th May 2015 at: https://www.uscg.mil/auxiliary/training/tct/chap5.pdf

University of Sheffield. (2015). SPSS for data processing [online]. Last accessed 12th August 2015 at: http://www.shef.ac.uk/lets/strategy/resources/evaluate/general/data-analysis/spss

Vis, F. (2013). A critical reflection on Big Data: Considering APIs, researchers and tools as data makers [online]. Last accessed 24th August 2015 at: http://journals.uic.edu/ojs/index.php/fm/article/view/4878/3755

Waddington, D. (2012). A 'kinder blue': analysing the police management of the Sheffield anti-'Lib Dem' protest of March 2011. *Policing and Society: An International Journal of Research and Policy,* ***23***(1), 46-64.

Waddington, D. (2011a). Policing Contemporary Political Protest: From Strategic Incapacitation to Strategic Facilitation? [online]. Last accessed 24th August 2015 at: https://www.shu.ac.uk/research/cresr/sites/shu.ac.uk/files/190111-waddington-political-protest.pdf

Waddington, D. (2011b). Public Order Policing in its Contexts: From the 1980s to the Present Day. Last accessed 24th August 2015 at: http://www.sipr.ac.uk/downloads/Public_Order/Waddington.pps

Waddington, D. (2011c). Public order policing in South Yorkshire, 1984–2011: the case for a permissive approach to crowd control. *Contemporary Social Science: Journal of the Academy of Social Science,* ***6(***3), 309-324.

Waddington, D. (2007) Public Order Policing: Theory and Practice. *Willan*.

Waddington, D. (1992) Contemporary Issues in Public Order: A Comparative and Historical Approach. *Routledge*.

Waddington, D., Jones, K., and Critcher, C. (1989). Flashpoints: Studies in Disorder. *Routledge*.

Wakefield, A. and Fleming, J. (2009). The Sage Dictionary of Policing. London: Sage Publication

Walsh, A., and Hemmens, C.T. (2011). *Introduction to Criminology: A Text/Reader (SAGE Text/Reader Series in Criminology and Criminal Justice)*. Second Edition. SAGE Publications, Inc.

Wall, D. (1999). *Earth First! and the Anti-Roads Movement*. First Edition. Routledge.

Wang, D., Marshall, J., and Huang, C. (2016). Theme-relevant truth discovery on twitter: An estimation theoretic approach. In *Tenth International AAAI Conference on Web and Social Media*.

Watts, S. (2013). *Can big data help predict crime?* [online]. Last accessed 15th February 2017 at: https://www.bbc.co.uk/news/av/technology-22008497/could-big-data-help-the-police-predict-crime

weareFLINT. (2016). UK Social Media Demographics 2016 [online]. Last accessed 15th February 2017 at: https://www.slideshare.net/weareflint/uk-social-media-demographics-2016

Weller, Katrin, et al. (2014). Twitter and society. New York : Peter Lang.

Weller, K., et al. (2014). *Twitter and Society.* Publisher: Peter Lang. ISBN: 978-1-4331-2169-2.

Wijermans, N. (2011). Understanding crowd behaviour: simulating situated individuals [online]. Last accessed 24th August 2015 at: https://www.rug.nl/research/portal/files/14565243/13complete.pdf

Wirth, R. (2000). CRISP-DM: Towards a standard process model for data mining [online]. Last accessed 12th August 2015 at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.198.5133&rep=rep1&type=pdf

Witten, I.H. (2011). *Data Mining: Practical Machine Learning Tools and Techniques.* Third Edition (The Morgan Kaufmann Series in Data Management Systems). 3rd Edition. Morgan Kaufmann.

Woodside, A.G. (2010). *Case Study Research: Theory, Methods and Practice*. Emerald Group Publishing Limited, UK.

Wyllie, D. (2013). Investigating Twitter: Mining social media for intel [online]. Last accessed 12th August 2015 at: http://www.policeone.com/police-products/investigation/articles/6241241-Investigating-Twitter-Mining-social-media-for-intel/

Xu, J., Zhu, X., and Bellmore, A. (2012) Fast learning for sentiment analysis on bullying. In: *Proceeding of International Workshop on Issues of Sentiment Discovery and Opinion Mining*.

Yarow, J. (2013). TWITTER'S IPO FILING IS OUT! [online]. Last accessed 12th August 2015 at: http://www.businessinsider.com/twitter-ipo-filing-2013-10?IR=T

YouTube. (2017). Statistics [online]. Last accessed 12th March 2018 at: https://www.youtube.com/yt/press/statistics.html

Zhang, D., Wang, J., and Zhao, X. (2015). Estimating the Uncertainty of Average F1 Scores. *Proceedings of the 2015 International Conference on Theory of Information Retrieval - ICTIR '15*.

Zou, L., and Song, W. (2016). LDA-TM: A two-step approach to Twitter topic data clustering. *2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*.

10 Appendices

10.9 Publication

Towards a Social Media Research Methodology: Defining Approaches and Ethical Concerns

James Baldwin,
Department of Computing,
C3RI, Sheffield Hallam University,
Sheffield, United Kingdom,
email: j.baldwin@shu.ac.uk

Dr Teresa Brunsdon &
Dr Jotham Gaudoin,
Department of Engineering and Mathematics, Sheffield Hallam University, United Kingdom
email: t.m.brunsdon@shu.ac.uk
j.gaudoin@shu.ac.uk

Dr Laurence Hirsch,
Department of Computing,
Sheffield Hallam University,
United Kingdom
email: l.hirsch@shu.ac.uk

***Abstract*— Social media research and suitable methodologies and ethical approaches for analysing social media data are still emerging. This paper presents a methodology for projects using social media data alongside consideration of ethics within the social media analysis context. Earlier stages of the methodology will be expanded to develop a strategy for examining ethics alongside consideration of the relevant analysis techniques that may be employed. This will provide a comprehensive methodology that will provide a springboard for the clear and ethically sound scrutiny of social media data. We aim to present the challenges of using social media data, while the inclusion of ethical and legal aspects in this paper aim to draw researchers' attention to the peculiarity issues involved with dealing with social media data.**

Keywords—social media; methodology; strategy; methods; ethics; legal; lifecycle.

I. INTRODUCTION

Since 2011, interest has grown in social media from both the academic and industrial perspectives [1]. For example, Law Enforcement Agencies substantially increased their usage of social media data, with policy changes being implemented to adapt to social media and its possible uses after the 2011 London riots occurred [2][3]. This interest has to some extent been driven by the rapid increase in usage of social media networks and of internet accessibility; the internet was used daily or almost daily by 82% (41.8 million) of UK adults, compared with 78% (39.3 million) in 2015 and 35% (16.2 million) in 2006 [4]. Organisations now have social media teams to monitor events and actively release information, quickly reacting to situations of widespread interest [1]. A great deal of research both has helped to shape the future of social media research, but this remains in its infancy. Examples of this inside the UK include the Government Social Researchers [1], a research team within the UK government "ensuring ministers and policy makers have the data to understand social issues [5] and evaluating the policy responses to them", the Economic and Social Research Council, Ipsos MORI [6] and the Centre for Analysis of Social Media - part of a cross-party charity-run think tank DEMOS [7]. Outside the UK there are such things the Big Boulder Initiative [8] located in the United States, which markets itself as the "first trade association for the social data industry" and European Citizen Science Association in Europe that is looks to "connect citizens and science through fostering active participation" whether that is using social media or other platforms [9].

The Big Data characteristics of social media data as regards their volume, velocity and scope has created a need for methodological innovations that are suited towards investigating social media data and their overall lifecycle and which apply both qualitative and quantitative approaches [10]. Quantitative methods seem to be the most popular in research to date, but analyses are certainly not restricted to this approach [10]. For example, new approaches in qualitative research are being formed in areas ranging from narrative analysis, to so-called *thick* data that document human behaviour and the context of that behaviour, to the analysis of non-verbal data such as sound and images, to combining and linking data - both text and interactions - from different platforms across times and contexts. Given this vast, expanding area of research, scholars will need to acquire new skills to explore, analyse and visualise their findings and situate them into their appropriate contexts [10], and will also need to be able to make appropriate ethical considerations for their research.

There is a need for further development of a clear methodology drawing together the already extant building blocks of good practice displayed both in researchers' papers [11] and by organisations, such as Canadian's government "Social Media Data Stewardship" (SMDS) project, that reduce the bias and flaws in social media data analysis. SMDS focuses on the data management processes applied in the context of using social media data. In the methodology section, we will discuss difficulties that are encountered when trying to find a social media lifecycle that has a clear defined strategy from start to finish, as without a clear approach to follow, such research can be a difficult experience for scholars embarking on work in this field.

The ethical perspective of extracting and collecting social media data in particular demands further consideration [10]. This is very important as it ensures the public's data are protected and are represented in a fair and respectful manner, whereby a tweet or post is not been taken out of context or used inappropriately. Ethics must be taken into consideration when going through each stage of the methodological framework. This paper will focus on the social media research methodology process,

while simultaneously considering the relevant ethical concerns.

The sections to be covered in this paper will be in accordance with the social media project lifecycle presented in Section 2. This will look to build upon existing methodological frameworks for social media research and, in particular, the GSR's social media lifecycle. This lifecycle was not originally designed for research purposes, and so must be modified to be fit for such a purpose, but it will be seen that it provides us with a good starting point from which to begin. Other approaches will also be considered and these will be merged in order to create a hybridised lifecycle that forms the essence of the methodology presented here. In Section 3, we will discuss the ethical concerns that can impact the social media research strategy and its lifecycle. In Section 4, conclusions will be drawn from the paper.

II. METHODOLOGY

Section 2 discusses a series of social media research strategies and how they are integrated into our social media lifecycle.

A. Social Media Research Strategy

Upon reviewing a wide range of papers, it was noted [11]-[13] that some provided an excellent, thorough description of the steps they took in their research. However, it was often found that the initial stages of the research that would be needed for a complete addressing of any research question were poorly defined. The available literature tends to be project specific in its approach and is therefore not immediately suitable for generalisation to other research - not unexpected, given that social media research methodology is a topic still in its infancy. From an early researcher's standpoint in particular, it may be difficult to know where to start in the area and to identify what decisions need to be taken to form a social media methodology for the project in question.

The research community and other organisations are trying to come up with better ways to express their social media strategies, such as the SMDS project, which *"focuses on studying practices behind and attitudes towards the collection, storage, use, reuse, analysis, publishing and preservation of social media data"* [14]. SMDS has produced a social media data process that aims to clarify for researchers the layout and order of each phase that may be required in a social media data project. SMDS focuses on the data management process of social media data and aims to help researchers to consider their attitudes towards the data they wish to work with [14]. What we aim to do in this paper is to identify a *complete* set of stages for any social media research project lifecycle to follow, including within this the SMDS insights into data management, as these touch on highly pertinent points within the overall process.

Having found the nascent SMDS data management paradigm, we continued the search for a full social media project lifecycle. While this proved impossible to source as no such lifecycle yet exists, we did encounter a somewhat developed social media research project lifecycle created by the UK Government Social Research (GSR) service. The GSR based its lifecycle on the Cabinet Office framework for data science projects, as it had *"numerous parallels here"* [1, p8]. This lifecycle has been tested on two social media projects within Government, namely, using Twitter to predict cases of Norovirus and assessing the experiences of the 20th Commonwealth games held in Glasgow, producing reports on the analysis of broadcast and online coverage. There is no publically available information on whether or not this social media lifecycle was in fact a success. However, GSR produced outcomes that may be a measure for potential successes. For example, the Commonwealth games on Twitter were in the top 10 highest sporting event hashtags of the year, generating a highly positive contribution to Scotland and Glasgow both internationally and within the rest of the UK [15]. Furthermore, GSR identified that between 14/06/14 to 06/08/14, there were 3.2 million mentions of the Commonwealth Games on social media in the English language. There were other positive outcomes, but what this allows GSR to do is to identify where future improvements can be made with the organisers in raising the profile for relevant cities and events [1] [15]. In the sequel, we shall aim to integrate aspects of the GSR service lifecycle and the SMDS data management process alongside our own insights into the social media project lifecycle.

B. Our integrated social media project lifecycle

The GSR social media project lifecycle [1] consists of seven stages: Stage 1: Rationale – Business/Citizen Need, Stage 2: Data, Stage 3: Tools and Output, Stage 4: Research Phase, Stage 5: Implementation/Publication/Action, Stage 6: Evaluation and finally Stage 7: Business as Usual. While this is a useful basic framework that will help to guide researchers through their social media projects, it still requires further development and refinement as the considerations outlined at each stage are given in little detail. Furthermore, this lifecycle is applied in a commercial and governmental context, which can make it difficult to know what to do at each step from a research perspective. Nevertheless, we have chosen to adopt this framework as a starting point as it proved itself helpful in structuring our own initial social media research project. The research we are conducting aims to enhance the analysis of social media in the context of public (dis-)order events. This investigates how social media data are stored (big data issues), collected, analysed (text mining and sentiment analysis) and then disseminated (to the police, to help predict when disorder may occur). This will form part of the creation of a model to analyse social media data to try to predict the escalation of such events and our research is presently ongoing. We will adapt the GSR lifecycle to suit the needs, aims and goals of research projects (as opposed to governmental projects), and a diagram showing the relevant adaptations is displayed in Figure 1.

Figure 1. Social media research project lifecycle

The steps in the lifecycle are explained below. We will outline the purpose of each step and show where modifications have been made to the GSR lifecycle. The lifecycle explained below will be informed by the pilot study we conducted, which has involved analysing Twitter data around the time of the Baltimore riots, with the aim of developing models to identify potential riots before they occur.

1) In [1], stage 1 (Rationale – Business/ Citizen Need) is described as a need to think about social media's attributes (e.g. speed, cost, real-time production). On the basis of these attributes, there are suggestions for the business or citizen's need to be based on: *"using insight to deliver a more timely service to the citizen with fewer resources through the support of social media analysis than would have been possible with traditional means."* [1, p9]. To measure if the project is delivering a timely and resource efficient service to the citizen can be difficult to determine in some cases without actually conducting the project. A rationale for the research must be established, as without this the project will likely lack focus and be too broad, weakening any results or insights obtained. This means that valuable resource that could potentially be better utilised elsewhere is being wasted. While nothing new has been added to this section compared to the GSR lifecycle, we have placed into the appropriate research context. This stage in our process is important, as one must have a question to drive the collection and analysis of data in research and, as outlined by [10], one should not let the data drive the researcher. Without a suitable research question, the project would lack purpose. The rationale for the project we carried is outlined above.

2) Stage 2 is a new step which has been introduced called "Selection of Potential Method(s)s". This step is required to help adapt this commercial lifecycle into a research context where consideration must be given as to which methods (for example, case study or archival research) will be applied in the research process. This must be decided early on in the process, so that the following stages can take this into account when making relevant decisions in the latter phases of the lifecycle. If this step is not undertaken explicitly in a research context then results may be obtained that are of a particular nature, without account having been taken of the fact that the nature of the methods employed is inextricably linked with one's research outputs. This may cause a loss of momentum in the stages ahead, where special account would have to be made for the method or methods employed. For our particular research, we selected a case study-based approach to allow us to work with particular disorder events immediately and then attempt to generalise these to the wider public order context.

3) "Data" is now stage 3 of the lifecycle. In [1, p9] it is emphasised that *"The primary purpose of this data is not for research so consideration should be given to representativeness, robustness and ethics."* This statement is confusing, as the same level of rigour would apply in a research context. In this section, the researcher must justify the datasets to be used in the project and examine any necessary ethical considerations regarding the use of the social media data in question in their research. The original purpose of this section remains the same as in the original GSR lifecycle. This phase considers which dataset(s) may be explored to answer the research questions of the project. There is extra emphasis on selecting the correct data as cost may well be an issue here, more so than for a government entity, depending on the size of dataset required for the research, given the finite nature of research grants in particular. This step is also useful in providing time to think carefully about the selection of data-sets. If the data are chosen without due care then this will impact the cleaning, analysis and output of the project, though given the emerging nature of social media technology, it can of course be difficult to fully understand the range of data and metadata that are available before one already has a sample to hand. To that end, collection of a small pre-sample of data can also be a useful initial substage here. The dataset used for the pilot study is based on collecting live data from the 2015 Baltimore riots, USA. This pilot study will help to inform the collection of further datasets, on which the pre-processing and data manipulation scripts developed for the Baltimore data can be re-run.

4) Stage 4, "Tools and Outputs" is named the same as in the original GSR lifecycle. In this phase, the use of specialised social media tools can help to make cleaning and analysis of the collected data easier for researchers. Furthermore, social media data may require manipulation to *"render it useful in a social research setting"* [1, p9]. The outputs from analysis of these data can range from traditional reports showing present findings to predictive models designed to solve real time problems. GSR's process for this step is kept, but

in addition to this, the researcher must outline their data collection strategy to show how relevant data in relation to any research questions will be obtained, as well as considering how those data will be stored and whether single or multiple platforms are to be used as this will have an effect on the tools chosen. There are a plethora of tools available for data acquisition, processing and analysis and the tools to be used must be selected with care to ensure that they are both suitably secure and efficacious for the data in question, otherwise, time will be invested in tools that are not appropriate for large scale data retrieval (not all return the same metadata, for example), cleaning and/or analysis. The tools selected will depend upon the platform from which data are to be extracted. In our case, since we are dealing with Twitter data we chose NVivo NCapture to extract a live sample of data from the Baltimore riots and used R for data manipulation. For the retrospective datasets that we collect in the future, we will instead be using DiscoverText for acquisition. This tool is widely used in the research community because it provides access to one of the cheapest ways to retrieve a complete historical record from Twitter's official provider GNIP. Even though the extraction and analytical tools are being selected at this stage, the actual techniques for analysis will be investigated in stage 5.

5) Stage 5 was originally named "Research Phase" in the GSR lifecycle [1], rather than "Analysis". Clearly, given that we are aiming to develop a full research lifecycle, the former name is no longer appropriate. This step emphasises that care must be taken regarding the representativeness of data to mitigate any bias in the analysis. Lastly, *"Care should be taken to ensure research generates a dataset of a size which can be handled by the subsequent analytics programs."* [1, p10]. This is an important aspect to consider, as the volume of data produced can be on a very large scale. This could break the confines of some analytical programs' constraints. Other Big Data characteristics (namely: variety, veracity, velocity and virtue) and the type of techniques applied by the researcher can have an influence on the choice of analytical tool adopted to achieve their aim(s) [10]. The naming of this section has been selected to align with its focus on preparing the data for the analysis, helping to identify whether the chosen analytical tools need to be changed to handle the dataset(s) in question and to establish which techniques (in our case, change point identification, sentiment analysis and machine learning) should be applied to analyse the data to assist in responding to a research aim and answering relevant research questions. The selection of techniques to analyse the data is a complex process that is dependent on the investigators' level of experience of the techniques in question while also ensuring that they will suit the dataset(s) chosen. For example, in our pilot study, the selection of sentiment analysis techniques for a newcomer to a developing field can be fraught with difficulties as different papers suggest different techniques to use and most do not provide a concrete path to understanding the basics before choosing what path to follow. Social media analysis is a developing area and at present one does wonder if the techniques available are effective enough for any given specific domain, whereas in other fields techniques may well have been tried and tested over many years. In our experience within the pilot study, this led to it taking a considerable length of time to make a decision, which is why it's appropriate for this consideration to have a stage of its own. Another consideration to make at this stage is whether the researcher has the appropriate equipment to process Big Data and explore the intricacies of the dataset chosen using the desired tools. For example, initially within our research, using the R language presented some issues when processing a large amount of data, as R Studio is single threaded. This meant the PC being used was inadequate and required an upgrade due to poor single threading performance. An assessment must be made early on as to whether the PC or Cloud selection has the processing power to analyse the data in a reasonable amount of time (or indeed at all if there are memory considerations).

6) Stage 6 was originally entitled "Implementation/ Publication/Action" and has been renamed to "Implementation" here. In [1], it is originally emphasised that social media research is in its infant stages and that the likelihood is that the work being carried out will be exploratory. Any successful *"outcome or otherwise should be communicated"* [1, p10] to the interested communities to build on this in future work, which is the same in business as in research. To assist in these steps the researcher can include the good practice from the SMDS approach on "publishing" to "reuse/sharing" and "preservation" [14]. Publication is one of the steps in this section as dissemination of research is clearly vital. The GSR lifecycle emphasises successful outcomes, but as this is now named "Implementation", there is a new focus, more appropriate for research, on making sure the project requirements and specifications as previously outlined above are implemented in practice so as to achieve the aims of the project. For example, in this step we extracted the data with NVivo NCapture, cleaned them and analysed them to detect the sentiment within each Tweet and identify significant changes of sentiment within the timeframe over which the data were collected by using R. It was appropriate that this all took place within this phase, as one step flowed to the next with purpose and direction to contribute to the aim of the project. In addition, to this, ethical consideration must be given further thought at this phase to how any data are shared and preserved, but this data management process will not be discussed in this paper, as we shall

focus on the legal and ethical considerations of social media data usage, which will look in particular at publication dilemmas. Publication is included in the last phase of the lifecycle instead as we must implement and (in particular) evaluate *before* we can publish within the research context. In our own context, had we attempted to include publication here alongside analysis, this stage would have become confused by the lack of evaluation. Furthermore, given the paucity of the quality of social media data, we required additional focus on relevant cleaning of the data and attempting to consider publishing at the same time would have resulted in a loss of momentum.

7) Stage 7 (Evaluation) is included in the lifecycle due to the immaturity of social media research compared with other more established research fields. In [1], there is a focus on the evaluation of exploring what value there is in social media research compared to traditional methods. It suggests that this stage will confirm whether not social media was specifically required "to respond to a business or citizen need" [1, p10]. This stage will remain the same as outlined in GSR's lifecycle but with a rather different focus. Where the GSR strategy considers whether or not there was value in the use of social media data, the researcher's focus will be on how effective the use of such data was in addressing the research aims and questions. A stage devoted to evaluation is important, as through evaluation we can identify whether our techniques have been effective in answering any research questions. For example, in our case, we aim to consider whether using a lexicon dictionary approach over machine learning for detecting sentiment provides a greater level of accuracy within the framework we have set. We have not yet completed this section of the lifecycle for our own work on social (dis-)order, but this stage of the pilot study has shown us which techniques are less effective (e.g. Latent Dirichlet Allocation) for this specific study and allowed us to apply a greater focus on others (e.g. Changepoint identification).

8) Stage 8 has been renamed from "Business as Usual" as it is in the GSR lifecycle [1] to "Knowledge Management" in order to fit the research context. The original purpose of this phase remains, but with the addition of publication to emphasise its importance in this context. This phase re-evaluates research techniques in order keep research up-to-date with any modern research techniques and to think how about how any knowledge gained about social media research methods themselves can be transferred to others to instil good practice. This stage can be commenced once a significant part of the cycle is completed. Publications are crucial way of sharing good practice within the research community and can then lead to subsequent further research after interactions with the community, leading us back to stage 1 to begin a new project and frame suitable new research questions. The pilot study's outcome has informed us that this original lifecycle with a series of changes can be placed into a research context that is effective in guiding social media projects. These findings will be shared in the form of publications and with other researchers through other means of communication such as conferences.

It is important to note the lifecycle is not only to be used as a single iteration. A researcher can go repeat stages to develop the project through one or many iterations. Furthermore, this lifecycle itself will be further evaluated when cycling through it again within the rest of our research project. Having outlined a possible lifecycle for social media research, in the next section, we discuss the ethical and legal considerations that must be made throughout the social media research lifecycle.

III. ETHICAL AND LEGAL CONSIDERATIONS

Technological advancements are outpacing developments in research governance and what is agreed as good practice. The ethical code of conduct that we rely on for guidance for collection, analysis and representation of data in this digital era is not up-to-date [16][17]. Social media is ethically challenging because of its openness in relation to the availability of data. The Terms and Conditions of these platforms (including Twitter, Facebook, YouTube, Weibo, Qzone, Reddit, LinkedIn and other global social media platforms) state that users' data is available for third parties, so in accepting these, users are giving legal consent for their data to be made available [18]. As [19] outlines *"Just because it is accessible doesn't mean using is ethical"*, which means that researchers must evaluate their positions carefully, as to whether using the data is or is not ethically sound.

Datasets with this scale of social interaction, speed of generation and level of access are unprecedented in the social sciences. This has led to many published papers that include complete tweets and/or usernames without informed consent [18]. This seems to have happened because of the openness of some social media platforms, thus leading to assumptions that these are 'public data' and that projects using such data therefore do not require the same level of scrutiny by an ethics panel as do studies using data collected by more standard methods, such as interview or questionnaires [18]. Some universities may have not caught up with the pace of technology and this is often reflected in their ethical policies and within their forms dealing with ethical considerations. Even where ethics panels have already scrutinised such data, they may still deem it to be 'public data' due to the lack of a suitable framework to evaluate the potential harm faced by those whose ostensibly public data is used in the research in question [17]. In some cases, ethical approval is not required per se, but it is suggested by a given university's policy that researchers consult resources, such as the Association of Internet Researchers, that can help to ensure that any social media data are used in an ethical fashion [17] [20].

Despite noting above that some ethical panels are not making much consideration about the ethical use of social media data, there is some evidence to suggest that a

number of universities are making strides towards updating their ethical guidelines with regards to social media data. As one such example, the University of Sheffield has a research ethics policy note that raises many important points that can be considered in other institutions [21]. This note indicates that research must have ethical approval *before* a dataset can be extracted. However, this may pose both a financial and a contemporaneity problem. If the researcher wants to use historical data that will in any case come at a cost then this will be the case with or without prior ethical authorisation. However, if the data cannot be extracted on-the-fly because ethical approval is taking time to obtain, then the institution's budget would have to be prepared to pay for those data in the long term. Furthermore, if the researcher is considering topics of current interest and wishes to amend their search criteria as data come in, it may not in fact even be *possible* to seek suitable a priori approval. Of course, planning in advance is well advised here, but there are times when one cannot predict the topics of research interest that will arise today, tomorrow or in many weeks' time, which makes it difficult to plan such requests in advance. This policy is thought provoking, as it makes the researcher think about the importance of ethics in the very early stages of their research and the requirement for ethical approval for social media research is clearly a step in the right direction towards ensuring high ethical standards. However, as noted above, it may be financial unviable, or prevent the collection of data required for some projects. To that end, we would recommend that perhaps there be a fast track ethical approval system for time-critical social media data projects so that on the one hand they receive suitable ethical scrutiny, while on the other they can also proceed in a timely manner, enabling researchers to react to current events of public interest.

According to a series of survey findings from [22] and [23], it appears that there is a disconnect between the practices of researchers in publishing content on social media posts and *"users' views of the fair use (includes accuracy) of their online communications in publications and their rights as research subjects."* [17]. The decision-making process in one's ethical approach to social media data must consider the expectations of social media users as regards their personal privacy. In addition to this, the researcher must review the nature of the information from a user on social media alongside its originally intended purpose.

Users on social media *"may not intend for their data to be used for their [researchers'] purposes"* [24] and have, therefore, not consented to it being used for research. Considerations must be given to possible risks to the users whose data are being employed in any research. We must recognise that social media research transcends the usual boundaries of geography and standard methodologies. This means that a scholar's research design must ensure that it satisfies the legal regulations and terms of service of each platform as well as those platforms' hosting countries' laws and the laws where the researchers are based. This also includes institutional guidelines, the privacy and expectations of users and their vulnerability from publications covering their activities, the reuse and publication of data and how users' contributions are anonymised [24]. The application of ethics must consider the concerns raised above. If researchers and organisations are not careful in their approach, the disconnect between researchers and users may grow further. A lack of action regarding such ethics could lead to a series of undesirable consequences, such as users calling on social media platforms for changes in their terms of service to restrict the use of their data. The impact of this may make it extremely difficult to use social media data for research designed for the public good.

Social media research ethics as specified above requires further development and awareness to ensure that the public's data are represented in their context in an accurate, respectful and fair way [10] [25]. Ethics of social media data analysis is of significant importance and is hotly debated in the research community (by organisations such as, the Social Research Association [26], the Academy of Social Sciences [27], and the New Social Media, New Social Science [28]) and outside of it, where improvements are continually being made to relevant ethical frameworks [10]. Ethics could be applied in the sense of one's own morality and standard of ethics, but the problem with this is that not everyone may have the same high ethical standards. Indeed, one may think that they have a high set of standards when in actuality their standards are lower than they believe and overall this is a slippery slope as it is open to suggestions of improper usage as there is no conformity to an agreed set of rules.

Current ethical guidelines are an ongoing area of development amongst research institutes and other organisations. There are a series of organisations that have produced a set of guidelines to follow, all of which support a high standard of ethical practice in social media research. Some examples of these organisations and efforts are provided below.

- A Canada Research Chair has emerged from a five-year partnership with SMDS. This project aims to address the concerns of incoherent and inadequate practice in social media research and suggests a set of guidelines for conducting large scale and aggregated analysis through social listening [14] on sensitive topics, such as medical and religious data [29].
- Ipsos MORI (funded by institutes such as the EPSRC, ESRC, CASM and DEMOS) is a market research organisation in the UK that is "curious about people, markets, brands and society", where they "deliver information and analyses by making it faster and easier to navigate our complex world and aid clients in making better decisions." IPSOS MORI produced a guide that examines and reviews the ethical, legal and regulatory framework for embedding ethics in social media research [30]. When considered alongside the SMDS framework, these provide a comprehensive set of user-driven principles to help manage all aspects of social media data in research, such as how to decide and handle the use of reproduced tweets - especially those that concern sensitive topics.

- The Economic and Social Research Council (ESRC) has an "ESRC Framework for research ethics", which contains a few social media guidelines [20] that can be put into practice. As social media ethics develops, we would suggest that the ESRC might wish to consider the addition of further guidance aimed towards helping social media researchers, particularly newcomers to the field, to navigate the uncertainty and confusion of this nascent field to help to ensure that they meet a high standard of ethics.
- The Government Social Research (GSR) team used a data science framework and incorporated a social media element into this directly. This report shows some pertinent core principles for the researcher that must be considered when conducting any social media research [1]. There are many important ethical considerations given, such as "*Core principle 4: Avoidance of personal and social harm"* [1, p20] and *"Core principle 5: Non-disclosure of identity"* [1, p20] which are straightforward and clear to understand.

The above ethical guidelines cover different areas, for example, the SDMS guidelines are focused on the actual conduct of social media analysis, the IPSOS MORI framework covers legal and regulatory issues, the ESRC guidelines are rather generic and do not yet constitute a concrete approach while the GSR team have simply appended to their current ethical framework a social media element, so that the framework is more specialised towards social media [25]. There are calls from [25] for institutions' ethics committees to integrate requirements into the approvals documentation (by specifying which ethical guidelines would be applied in one's research); as [25] suggests there is a low level of ethical awareness amongst researchers applying social media data mining in their studies.

Now, all these guidelines provide very important points, but their multiplicity creates difficulties for the researcher as there are still uncertainties around the ethics of social media research in part because these guidelines do not always agree. Of course, this area is still in its early stages of development. In addition, what makes this area even more difficult is the terms and conditions set by the individual social media platforms. These can be hard to interpret because of the legal terminology or may be otherwise ambiguous and different platforms have different terms of service, such that it can be difficult for multi-platform research to adhere to them all simultaneously. Moreover, the terms and conditions can create ethical concerns for publication - for example, Twitter will not allow tweets to be presented without usernames [10], which can make it difficult to protect participants from potential harm. If the data are highly sensitive and the username is published, then the effect of linking the user to these data and the research may cause an effect within the public sphere. For example, the subject may receive positive responses, thereby boosting their reputation, or, perhaps more seriously, highlighting negative tweets may damage the mental or physical wellbeing of those mentioned within them.

Online research poses a greater risk to upholding confidentiality than does protecting offline research [18]. One reason for this is that at present there is a permanent record of what has been posted online. For instance, any quotation used can lead directly back that user in question with the use of a search engine [18]. This raises concerns over the anonymity of data. For example, as noted above, Twitter's data sharing licensing policy allows the sharing of Tweet IDs only, to ensure the data collection process is reproducible [18]. Using the identification ID provides a way to obtain the same dataset from Twitter's API. These IDs are unique and are easily searchable on the web to locate each tweet. This can be a cause for concern as it makes it easier to de-anonymise the data, so if the data are highly sensitive then a choice has to made as to whether that ID should be excluded from being shared if it causes an ethical concern [18]. Furthermore, the anonymisation techniques we can apply now may become easier to deanonymise in the future due to technological advancement.

In the UK, we must also take other laws into consideration, such as the Data Protection Act 1998, as researchers need to comply fully with the data protection principles laid out therein. Section 33 of the Data Protection Act 1998 allows exemptions to be made in accordance with principles 2 and 5 of the act for personal data used in research [31]. Recent developments within the UK Government suggest they are looking to form a council of data ethics *"to address the growing legal and ethical challenges associated with balancing privacy, anonymisation of data, security and public benefit."* [32] and also to implement the General Data Protection Regulation on the 25th of May 2018 [33]. Researchers will have to take these developments into account in their future practice as it may impact their social media research. Even after Brexit, the General Data Protection Regulation (which includes similarities with the existing UK Data Protection Act 1998) will be adopted into UK law [33]. It is essential that researchers keep abreast of any legal developments and keep up-to-date with good practise in their relevant area so as to make the best possible ethical use of social media data.

The concerns outlined above regarding the ethical challenges of using social media data can make for a difficult challenge for the social media researcher. The best course of action the researcher community can take is to address concerns and difficulties on case-by-case basis, thereafter trying to update guidelines and frameworks to deal with such cases. Genuine mistakes might have been made in the research community, which both individual researchers and the community as a whole can learn from. If a researcher has made a genuine ethics-related mistake in their work and has demonstrated remorse, then we as a community need to forgive and look to further strengthen the ethical standards and frameworks available to us. Indeed, ethical concepts are not just hoops to jump through in the early phases of research, but concepts requiring ethical inquiry [18], which may in itself take time. Mistakes may not be recognised until well after they have occurred and numerous judgements are possible, which can provide uncertainty and ambiguity, but this is likely to apply to any research [18]. Ethical

considerations will be in a constant state of assessment throughout any project and each case that arises during the research process can be worked through using a set of context-specific decisions. In addition to this, researchers must be guided by core ethical principles set by their employing organisations and external bodies, while also employing an appropriate mixture of the frameworks as laid out above, to ensure that the highest ethical standards are followed in any research.

There is a need to improve ethical assessment and one way to do this is to create a value-based ethical culture and practices in the research community and within other organisations for the development and deployment of intelligent systems both within the UK and elsewhere. This is known as Value Based Design (VBD) [34]. To do this, one must identify, enhance and ultimately embrace management strategies and social processes that facilitate value-based ethics within their design process. This could be included as an additional step in the framework in a future development, as it may provide a way to ensure a higher standard of ethical practice in the future.

IV. Conclusion

This paper has taken an existing methodology, the GSR lifecycle, and created from it a new social media lifecycle suitable for the research context. This was illustrated via a new diagram (Figure 1) that contains steps adapted to incorporate changes that are required for use in a research context. Alongside this, a number of ethical concerns have been explored and we have highlighted a series of pertinent points to consider in any future social media research project. Overall, this paper has sought to provide an easier way for researchers to enter the domain of social media research and then conduct relevant research, while providing an insight into the importance of the relevant ethical considerations in this area. Future research directions could include widening the framework beyond the UK, to other domains such as the wider European Union, the United States and Canada in a more detailed fashion and further thought could be given to how to expand the framework to include VBD.

References

[1] "Using social media for social research: An introduction", *Assets.publishing.service.gov.uk*, 2016. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/524750/GSR_Social_Media_Research_Guidance_-_Using_social_media_for_social_research.pdf. [Accessed: 09- Aug- 2018].

[2] J. Crump, "What Are the Police Doing on Twitter? Social Media, the Police and the Public", *Policy & Internet*, vol. 3, no. 4, pp. 1-27, 2011.

[3] M. Downes, "UK Police Forces using Social Media: Twitter, facebook, YouTube, Google+ and Hangouts On Air - The January 2013 Survey by Mike Downes", *Whatsinkenilworth.com*, 2013. [Online]. Available: http://www.whatsinkenilworth.com/2013/01/uk-police-forces-using-social-media.html. [Accessed: 08- Aug- 2018].

[4] "Internet access – households and individuals, Great Britain - Office for National Statistics", *Ons.gov.uk*, 2016. [Online]. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/internetaccesshouseholdsandindividuals/2016. [Accessed: 08- Aug- 2018].

[5] "Government Social Research Service | Civil Service Fast Stream", *Faststream.gov.uk*, 2018. [Online]. Available: https://www.faststream.gov.uk/government-social-research-service. [Accessed: 22- Aug- 2018].

[6] "Global market and opinion research specialist | Ipsos MORI", *Ipsos.com*, 2018. [Online]. Available: https://www.ipsos.com/ipsos-mori/en-uk. [Accessed: 22- Aug- 2018].

[7] "Big Boulder Initiative, Big Boulder Conference Big Boulder Initiative", *Bbi.org*, 2018. [Online]. Available: http://www.bbi.org/. [Accessed: 08- Aug- 2018].

[8] "Demos", *Demos.co.uk*, 2018. [Online]. Available: https://www.demos.co.uk/. [Accessed: 22- Aug- 2018].

[9] "About us", *European Citizen Science Association (ECSA)*, 2018. [Online]. Available: https://ecsa.citizen-science.net/about-us. [Accessed: 08- Aug- 2018].

[10] A. Quan-Haase and L. Sloan, *The SAGE handbook of social media research methods*. Los Angeles: Sage Publications, pp 1-9, 2017.

[11] D. Ruths and J. Pfeffer, "Social media for large studies of behavior," *Science*, vol. 346, no. 6213, pp. 1063–1064, 2014.

[12] R. Procter, F. Vis and A. Voss, "Reading the riots on Twitter: methodological innovation for the analysis of big data", *International Journal of Social Research Methodology*, vol. 16, no. 3, pp. 197-214, 2013.

[13] Y. Theocharis, W. Lowe, J. van Deth and G. García-Albacete, "Using Twitter to mobilize protest action: online mobilization patterns and action repertoires in the Occupy Wall Street, Indignados, and Aganaktismenoi movements", *Information, Communication & Society*, vol. 18, no. 2, pp. 202-220, 2014.

[14] "About – Social Media Data Stewardship", *Socialmediadata.org*, 2017. [Online]. Available: http://socialmediadata.org/about/. [Accessed: 09- Aug- 2018].

[15] "Analysis of XX Commonwealth Games Host Broadcast Coverage, Online Media and Official Digital Channels", *Gov.scot*, 2015. [Online]. Available: https://www.gov.scot/Resource/0048/00482015.pdf. [Accessed: 09- Aug- 2018].

[16] M. Williams, P. Burnap and L. Sloan, "Crime Sensing with Big Data: The Affordances and Limitations of using Open Source Communications to Estimate Crime Patterns", *British Journal of Criminology*, vol. 57, no. 2, pp. Pages 320–340, 2016.

[17] "Ethics Resources – Social Data Science Lab", *Socialdatalab.net*, 2018. [Online]. Available: http://socialdatalab.net/ethics-resources. [Accessed: 08- Aug- 2018].

[18] K. Beninger. Social Media Users' Views on the Ethics of Social Media Research. In: A. Quan-Haase and L. Sloan. (eds.) *The SAGE handbook of social media research methods*. Los Angeles: Sage Publications, p.53–73, 2017.

[19] D. Boyd, "Privacy and Publicity in the Context of Big Data", *Danah.org*, 2010. [Online]. Available: http://www.danah.org/papers/talks/2010/WWW2010.html. [Accessed: 08- Aug- 2018].

[20] "ESRC Framework for research ethics Updated January 2015", *Esrc.ac.uk*, 2015. [Online]. Available: http://www.esrc.ac.uk/files/funding/guidance-for-applicants/esrc-framework-for-research-ethics-2015. [Accessed: 08- Aug- 2018].

[21] "The University of Sheffield Research Ethics Policy Note no. 14, Research Involving Social Media Data", *Sheffield.ac.uk*, 2017. [Online]. Available: https://www.sheffield.ac.uk/polopoly_fs/1.670954!/file/Research-Ethics-Policy-Note-14.pdf. [Accessed: 08- Aug- 2018].

[22] M. Williams, "Towards an ethical framework for using social media data in social research", *Socialdatalab.net*, 2015. [Online]. Available: http://socialdatalab.net/wp-content/uploads/2016/08/EthicsSM-SRA-Workshop.pdf. [Accessed: 07- Aug- 2018].

[23] K. Beninger, A. Fry, N. Jago, H. Lepps, L. Nass and H. Silvester, "NatCen Social Research, Research using Social Media; Users", *Natcen.ac.uk*, 2014. [Online]. Available: http://www.natcen.ac.uk/media/282288/p0639-research-using-social-media-report-final-190214.pdf. [Accessed: 07- Aug- 2018].

[24] A. Bruns, "Challenges in Social Media Research Ethics | Snurblog", *Snurb.info*, 2017. [Online]. Available: http://snurb.info/node/2227. [Accessed: 07- Aug- 2018].

[25] J. Taylor and C. Pagliari, "Mining social media data: How are research sponsors and researchers addressing the ethical challenges?", *Research Ethics*, vol. 14, no. 2, pp. 1-39, 2017.

[26] "The SRA | Home of the Social Research Community", *The-sra.org.uk*, 2018. [Online]. Available: http://the-sra.org.uk/. [Accessed: 03- Oct- 2018].
[27] "Academy of Social Sciences", *Acss.org.uk*, 2018. [Online]. Available: https://www.acss.org.uk/. [Accessed: 03- Oct- 2018].
[28] New Social Media New Social Science, "#NSMNSS", *Nsmnss.blogspot.com*, 2018. [Online]. Available: http://nsmnss.blogspot.com/. [Accessed: 03- Oct- 2018].
[29] E. Yom-Tov, D. Borsa, I. Cox and R. McKendry, "Detecting Disease Outbreaks in Mass Gatherings Using Internet Data", *Journal of Medical Internet Research*, vol. 16, no. 6, p. e154, 2014.
[30] H. Evans, S. Ginnis and J. Bartlett, *Ipsos.com*, 2015. [Online]. Available: https://www.ipsos.com/sites/default/files/migrations/en-uk/files/Assets/Docs/Publications/im-demos-social-ethics-in-social-media-research-summary.pdf. [Accessed: 07- Aug- 2018].
[31] W. Ahmed, "Ethical Challenges of Using Social Media Data In Research", *YouTube*, 2017. [Online]. Available: https://www.youtube.com/watch?v=VeFMqL4Hj60. [Accessed: 07- Aug- 2018].
[32] "Government agree to set up 'Council of Data Ethics", *UK Parliament*, 2016. [Online]. Available: https://www.parliament.uk/business/committees/committees-a-z/commons-select/science-and-technology-committee/news-parliament-2015/big-data-dilemma-government-response-15-16/. [Accessed: 07- Aug- 2018].
[33] "Overview of the General Data Protection Regulation (GDPR)", *Ico.org.uk*, 2017. [Online]. Available: https://ico.org.uk/media/for-organisations/data-protection-reform/overview-of-the-gdpr-1-13.pdf. [Accessed: 05- Aug- 2018].
[34] "Methodologies to Guide Ethical Research and Design", *Standards.ieee.org*, 2018. [Online]. Available: https://standards.ieee.org/develop/indconn/ec/ead_methodologies_research_v2.pdf. [Accessed: 05- Aug- 2018].

10.10 Interrater agreement results

10.10.1 MR1 and MR2 results

2016 Anti-Austerity Inter Annotation Agreement Results				
Category	**1st rater**	**2nd rater**	**Percentage agreement (Tolerance=0)**	**Cohen's Kappa for 2 Raters (Weights: unweighted)**
Counted Positive	138.00	156.00	Subjects = 1500	Subjects = 1500
Counted Negative	715.00	200.00	Raters = 2	Raters = 2
Counted Neutral	647.00	1144.00	%-agree = 56	Kappa = 0.264
Total	1500.00	1500.00		z = 16.9
				p-value = 0
Positive Proportion	9.20	10.40	**Cohen's Kappa for 2 Raters (Weights: equal)**	**Cohen's Kappa for 2 Raters (Weights: squared)**
Negative Proportion	47.67	13.33	Subjects = 1500	Subjects = 1500
Neutral Proportion	43.13	76.27	Raters = 2	Raters = 2
Total	100.00	100.00	Kappa = 0.302	Kappa = 0.359
			z = 19.8	z = 17.3
Matched	840.00		p-value = 0	p-value = 0
Unmatched	660.00		**Krippendorff's alpha**	
Total	1500.00		Subjects = 1500	
			Raters = 2	
Proportion Matched	56.00		alpha = 0.271	
Proportion Unmatched	44.00			

Table 78 MR1 and MR2 Inter Agreement for Anti-Austerity

2016 Dover Inter Annonater Agreement Results				
Category	**1st rat**	**2nd rat**	**Percentage agreement (Tolerance=0)**	**Cohen's Kappa for 2 Raters (Weights: unweighted)**
Counted Positive	29.00	49.00	Subjects = 1500	Subjects = 1500
Counted Negative	1232.00	976.00	Raters = 2	Raters = 2
Counted Neutral	239.00	475.00	%-agree = 76.3	Kappa = 0.429
Total	1500.00	1500.00		z = 20.1
				p-value = 0
Positive Proportion	1.93	3.27	**Cohen's Kappa for 2 Raters (Weights: equal)**	**Cohen's Kappa for 2 Raters (Weights: squared)**
Negative Proportion	82.13	65.07	Subjects = 1500	Subjects = 1500
Neutral Proportion	15.93	31.67	Raters = 2	Raters = 2
Total	100.00	100.00	Kappa = 0.448	Kappa = 0.477
			z = 21.1	z = 20.2
Matched	1145.00		p-value = 0	p-value = 0
Unmatched	355.00		**Krippendorff's alpha**	
Total	1500.00		Subjects = 1500	
			Raters = 2	
Proportion Matched	76.33		alpha = 0.453	
Proportion Unmatched	23.67			

Table 79 MR1 and MR2 Inter Agreement for Dover

2016 MMM Inter Annoater Agreement Results				
Catego	1st rat	2nd rat	Percentage agreement (Tolerance=0)	Cohen's Kappa for 2 Raters (Weights: unweighted)
Counted Positive	127.00	178.00	Subjects = 1500	Subjects = 1500
Counted Negative	679.00	405.00	Raters = 2	Raters = 2
Counted Neutral	694.00	917.00	%-agree = 71.3	Kappa = 0.492
Total	1500.00	1500.00		z = 25.3
				p-value = 0
Positive Proportion	8.47	11.87	**Cohen's Kappa for 2 Raters (Weights: equal)**	**Cohen's Kappa for 2 Raters (Weights: squared)**
Negative Proportion	45.27	27.00	Subjects = 1500	Subjects = 1500
Neutral Proportion	46.27	61.13	Raters = 2	Raters = 2
Total	100.00	100.00	Kappa = 0.515	Kappa = 0.549
			z = 27.1	z = 22.6
Matched	1054.00		p-value = 0	p-value = 0
Unmatched	446.00		**Krippendorff's alpha**	
Total	1500.00		Subjects = 1500	
			Raters = 2	
Proportion Matched	70.27		alpha = 0.527	
Proportion Unmatched	29.73			

Table 80 MR1 and MR2 Inter Agreement for 2016 MMM

2015 MMM Inter Annonater Agreement Results				
Category	**1st rater**	**2nd rater**	**Percentage agreement (Tolerance=0)**	**Cohen's Kappa for 2 Raters (Weights: unweighted)**
Counted Positive	86.00	99.00	Subjects = 1500	Subjects = 1500
Counted Negative	835.00	535.00	Raters = 2	Raters = 2
Counted Neutral	579.00	866.00	%-agree = 70.3	Kappa = 0.5
Total	1500.00	1500.00		z = 24.3
				p-value = 0
Positive Proportion	5.73	6.60	**Cohen's Kappa for 2 Raters (Weights: equal)**	**Cohen's Kappa for 2 Raters (Weights: squared)**
Negative Proportion	55.67	35.67	Subjects = 1500	Subjects = 1500
Neutral Proportion	38.60	57.73	Raters = 2	Raters = 2
Total	100.00	100.00	Kappa = 0.516	Kappa = 0.54
			z = 25.5	z = 22.2
Matched	1069.00		p-value = 0	p-value = 0
Unmatched	431.00		**Krippendorff's alpha**	
Total	1500.00		Subjects = 1500	
			Raters = 2	
Proportion Matched	71.27		alpha = 0.511	
Proportion Unmatched	28.73			

Table 81 MR1 and MR2 Inter Agreement for 2015 MMM

10.10.2 MR1 and MR3 results

2015 MMM Inter Annonater Agreement Results				
Category	**1st rater**	**3rd rater**	**Percentage agreement (Tolerance=0)**	**Cohen's Kappa for 2 Raters (Weights: unweighted**
Counted Positive	86.00	43.00	Subjects = 1500	Subjects = 1500
Counted Negative	835.00	90.00	Raters = 2	Raters = 2
Counted Neutral	579.00	1367.00	%-agree = 43.1	Kappa = 0.0726
Total	1500.00	1500.00		z = 7.08
				p-value = 1.47e-12
Positive Proportion	5.73	2.87	**Cohen's Kappa for 2 Raters (Weights: equal)**	**Cohen's Kappa for 2 Raters (Weights: squared)**
Negative Proportion	55.67	6.00	Subjects = 1500	Subjects = 1500
Neutral Proportion	38.60	91.13	Raters = 2	Raters = 2
Total	100.00	100.00	Kappa = 0.0747	Kappa = 0.0785
			z = 7.34	z = 5.71
Matched	647.00		p-value = 2.12e-13	p-value = 1.14e-08
Unmatched	853.00		**Krippendorff's alpha**	
Total	1500.00		Subjects = 1500	
			Raters = 2	
Proportion Matched	43.13		alpha = -0.135	
Proportion Unmatched	56.87			

Table 82 MR1 and MR3 Inter Agreement for 2015 MMM

2016 MMM Inter Annoater Agreement Results				
Category	**1st rater**	**3rd rater**	**Percentage agreement (Tolerance=0)**	**Cohen's Kappa for 2 Raters (Weights: unweighted**
Counted Positive	127.00	12.00	Subjects = 1500	Subjects = 1500
Counted Negative	679.00	30.00	Raters = 2	Raters = 2
Counted Neutral	694.00	1458.00	%-agree = 48.9	Kappa = 0.0541
Total	1500.00	1500.00		z = 7.81
				p-value = 5.77e-15
Positive Proportion	8.47	0.80	**Cohen's Kappa for 2 Raters (Weights: equal)**	**Cohen's Kappa for 2 Raters (Weights: squared)**
Negative Proportion	45.27	2.00	Subjects = 1500	Subjects = 1500
Neutral Proportion	46.27	97.20	Raters = 2	Raters = 2
Total	100.00	100.00	Kappa = 0.0596	Kappa = 0.0704
			z = 8.51	z = 7.17
Matched	733.00		p-value = 0	p-value = 7.59e-13
Unmatched	767.00		**Krippendorff's alpha**	
Total	1500.00		Subjects = 1500	
			Raters = 2	
Proportion Matched	48.87		alpha = -0.0778	
Proportion Unmatched	51.13			

Table 83 MR1 and MR3 Inter Agreement for 2016 MMM

2016 Dover Inter Annonater Agreement Results				
Category	**1st rater**	**3rd rater**	**Percentage agreement (Tolerance=0)**	**Cohen's Kappa for 2 Raters (Weights: unweighte**
Counted Positive	29.00	19.00	Subjects = 1500	Subjects = 1500
Counted Negative	1232.00	91.00	Raters = 2	Raters = 2
Counted Neutral	239.00	1390.00	%-agree = 22.1	Kappa = 0.0294
Total	1500.00	1500.00		z = 5.07
				p-value = 3.94e-07
Positive Proportion	1.93	1.27	**Cohen's Kappa for 2 Raters (Weights: equal)**	**Cohen's Kappa for 2 Raters (Weights: squared)**
Negative Proportion	82.13	6.07	Subjects = 1500	Subjects = 1500
Neutral Proportion	15.93	92.67	Raters = 2	Raters = 2
Total	100.00	100.00	Kappa = 0.0342	Kappa = 0.0434
			z = 5.81	z = 5.94
Matched	332.00		p-value = 6.3e-09	p-value = 2.81e-09
Unmatched	1168.00		**Krippendorff's alpha**	
Total	1500.00		Subjects = 1500	
			Raters = 2	
Proportion Matched	22.13		alpha = -0.496	
Proportion Unmatched	77.87			

Table 84 MR1 and MR3 Inter Agreement for 2016 Dover

2016 Anti-Austerity Inter Annotation Agreement Results				
Category	**1st rater**	**3rd rater**	**Percentage agreement (Tolerance=0)**	**Cohen's Kappa for 2 Raters (Weights: unweighte**
Counted Positive	138.00	61.00	Subjects = 1500	Subjects = 1500
Counted Negative	715.00	41.00	Raters = 2	Raters = 2
Counted Neutral	647.00	1398.00	%-agree = 47	Kappa = 0.0881
Total	1500.00	1500.00		z = 10
				p-value = 0
Positive Proportion	9.20	4.07	**Cohen's Kappa for 2 Raters (Weights: equal)**	**Cohen's Kappa for 2 Raters (Weights: squared)**
Negative Proportion	47.67	2.73	Subjects = 1500	Subjects = 1500
Neutral Proportion	43.13	93.20	Raters = 2	Raters = 2
Total	100.00	100.00	Kappa = 0.119	Kappa = 0.175
			z = 13.9	z = 12.9
Matched	705.00		p-value = 0	p-value = 0
Unmatched	795.00		**Krippendorff's alpha**	
Total	1500.00		Subjects = 1500	
			Raters = 2	
Proportion Matched	47.00		alpha = 0.0111	
Proportion Unmatched	53.00			

Table 85 MR1 and MR3 Inter Agreement for 2016 Anti-Austerity

10.10.3 MR2 and MR3 results

2015 MMM Inter Annonater Agreement Results				
Category	**2nd rater**	**3rd rater**	**Percentage agreement (Tolerance=0)**	**Cohen's Kappa for 2 Raters (Weights: unweighted)**
Counted Positive	99.00	43.00	Subjects = 1500	Subjects = 1500
Counted Negative	535.00	90.00	Raters = 2	Raters = 2
Counted Neutral	866.00	1367.00	%-agree = 60.2	Kappa = 0.117
Total	1500.00	1500.00		z = 8.38
				p-value = 0
Positive Proportion	6.60	2.87	**Cohen's Kappa for 2 Raters (Weights: equal)**	**Cohen's Kappa for 2 Raters (Weights: squared)**
Negative Proportion	35.67	6.00	Subjects = 1500	Subjects = 1500
Neutral Proportion	57.73	91.13	Raters = 2	Raters = 2
Total	100.00	100.00	Kappa = 0.106	Kappa = 0.0874
			z = 7.88	z = 4.85
Matched	903.00		p-value = 3.11e-15	p-value = 1.24e-06
Unmatched	597.00		**Krippendorff's alpha**	
Total	1500.00		Subjects = 1500	
			Raters = 2	
Proportion Matched	60.20		alpha = 0.00849	
Proportion Unmatched	39.80			

Table 86 MR2 and MR3 Inter Agreement for 2015 MMM

2016 MMM Inter Annoater Agreement Results				
Category	**2nd rater**	**3rd rater**	**Percentage agreement (Tolerance=0)**	**Cohen's Kappa for 2 Raters (Weights: unweighted)**
Counted Positive	178.00	12.00	Subjects = 1500	Subjects = 1500
Counted Negative	405.00	30.00	Raters = 2	Raters = 2
Counted Neutral	917.00	1458.00	%-agree = 63.3	Kappa = 0.0804
Total	1500.00	1500.00		z = 9.09
				p-value = 0
Positive Proportion	11.87	0.80	**Cohen's Kappa for 2 Raters (Weights: equal)**	**Cohen's Kappa for 2 Raters (Weights: squared)**
Negative Proportion	27.00	2.00	Subjects = 1500	Subjects = 1500
Neutral Proportion	61.13	97.20	Raters = 2	Raters = 2
Total	100.00	100.00	Kappa = 0.0874	Kappa = 0.101
			z = 9.97	z = 8
Matched	949.00		p-value = 0	p-value = 1.33e-15
Unmatched	551.00		**Krippendorff's alpha**	
Total	1500.00		Subjects = 1500	
			Raters = 2	
Proportion Matched	63.27		alpha = 0.0737	
Proportion Unmatched	36.73			

Table 87 MR2 and MR3 Inter Agreement for 2016 MMM

2016 Dover Inter Annonater Agreement Results				
Category	**2nd rater**	**3rd rater**	**Percentage agreement (Tolerance=0)**	**Cohen's Kappa for 2 Raters (Weights: unweigl**
Counted Positive	49.00	19.00	Subjects = 1500	Subjects = 1500
Counted Negative	976.00	91.00	Raters = 2	Raters = 2
Counted Neutral	475.00	1390.00	%-agree = 37.2	Kappa = 0.058
Total	1500.00	1500.00		z = 6.6
				p-value = 4.12e-11
Positive Proportion	3.27	1.27	**Cohen's Kappa for 2 Raters (Weights: equa**	**Cohen's Kappa for 2 Raters (Weights: squared)**
Negative Proportion	65.07	6.07	Subjects = 1500	Subjects = 1500
Neutral Proportion	31.67	92.67	Raters = 2	Raters = 2
Total	100.00	100.00	Kappa = 0.0653	Kappa = 0.0794
			z = 7.4	z = 7.32
Matched	558.00		p-value = 1.33e-13	p-value = 2.41e-13
Unmatched	942.00		**Krippendorff's alpha**	
Total	1500.00		Subjects = 1500	
			Raters = 2	
Proportion Matched	37.20		alpha = -0.236	
Proportion Unmatched	62.80			

Table 88 MR2 and MR3 Inter Agreement for 2016 Dover

2016 Anti-Austerity Inter Annotation Agreement Results				
Category	**2nd rater**	**3rd rater**	**Percentage agreement (Tolerance=0)**	**Cohen's Kappa for 2 Raters (Weights: unweighte**
Counted Positive	156.00	61.00	Subjects = 1500	Subjects = 1500
Counted Negative	200.00	41.00	Raters = 2	Raters = 2
Counted Neutral	1144.00	1398.00	%-agree = 78.4	Kappa = 0.232
Total	1500.00	1500.00		z = 14.8
				p-value = 0
Positive Proportion	10.40	4.07	**Cohen's Kappa for 2 Raters (Weights: equal)**	**Cohen's Kappa for 2 Raters (Weights: squared)**
Negative Proportion	13.33	2.73	Subjects = 1500	Subjects = 1500
Neutral Proportion	76.27	93.20	Raters = 2	Raters = 2
Total	100.00	100.00	Kappa = 0.254	Kappa = 0.294
			z = 17.4	z = 13.8
Matched	1176.00		p-value = 0	p-value = 0
Unmatched	324.00		**Krippendorff's alpha**	
Total	1500.00		Subjects = 1500	
			Raters = 2	
Proportion Matched	78.40		alpha = 0.292	
Proportion Unmatched	21.60			

Table 89 MR2 and MR3 Inter Agreement for 2016 Anti-Austerity

10.10.4 MR1, MR2 and MR3 agreement

MMM 2015 Inter Annotation Agreement Results			Dover 2016 Inter Annotation Agreement Results	
Category of Agreement	**Agreed/Unagree**	**Proportion of Agreed/Unagree**	**Agreed/Unagreed**	**Proportion of Agreed/Unagreed**
All Agreed	571	38.07	280	18.67
J&G Agreed	498	33.20	865	57.67
J&S Agreed	76	5.07	52	3.47
G&S Agreed	332	22.13	278	18.53
Not Agreed	23	1.53	25	1.67
Total	**1500**	**100**	**1500**	**100**
MMM 2016 Inter Annotation Agreement Results			**AA 2016 Inter Annotation Agreement Results**	
All Agreed	634	42.27	626	41.73
J&G Agreed	420	28.00	214	14.27
J&S Agreed	99	6.60	79	5.27
G&S Agreed	315	21.00	550	36.67
Not Agreed	32	2.13	31	2.07
Total	**1500**	**100**	**1500**	**100**

Table 90 MR1, MR2 & MR3 Inter Agreement for each event

10.11 Breakdown for sentiment category: Precision and Recall Results

10.11.1 MR1 Results

10.11.1.1 Negative

Dictionaries	Precision	Recall	F1 Measu
sentiment_jockers_rinker	0.82	0.78	0.80
syuzhet_jockers	0.82	0.77	0.80
sentimentr_jockers	0.82	0.77	0.79
sentiment_vadar	0.84	0.62	0.71
syuzhet_afinn	0.82	0.62	0.71
combined_dictionary	0.77	0.59	0.67
sentimentr_huliu	0.86	0.54	0.66
sentiment_senticnet	0.72	0.59	0.65
syuzhet_bing	0.86	0.50	0.63
sentiment_loughran_mcdonald	0.80	0.52	0.63
sentimentr_sentiword	0.70	0.57	0.63
sentiment_berkeley	0.80	0.50	0.61
sentiment_inquirer	0.87	0.46	0.60
sentiment_stanford	0.61	0.45	0.52
sentiment_slangsd	0.59	0.42	0.49
sentiment_sentistrength	0.56	0.41	0.47
sentiment_nrc	0.84	0.26	0.40
syuzhet_nrc	0.87	0.26	0.40
sentiment_socal_google	0.74	0.20	0.31

Table 91 MMM 2015 Negative Precision and Recall Outcome (MR1)

Dictionaries	Precision	Recall	F1 Measu
syuzhet_jockers	0.73	0.67	0.70
sentiment_jockers_rinker	0.72	0.67	0.69
sentimentr_jockers	0.73	0.65	0.69
sentiment_vadar	0.80	0.57	0.66
syuzhet_afinn	0.77	0.56	0.65
sentiment_sentistrength	0.74	0.55	0.63
combined_dictionary	0.69	0.55	0.61
sentiment_senticnet	0.61	0.52	0.56
syuzhet_bing	0.78	0.43	0.56
sentimentr_huliu	0.76	0.44	0.55
sentiment_loughran_mcdonald	0.72	0.45	0.55
sentiment_berkeley	0.72	0.42	0.53
sentimentr_sentiword	0.56	0.48	0.52
sentiment_inquirer	0.75	0.37	0.49
sentiment_slangsd	0.49	0.48	0.49
sentiment_stanford	0.53	0.45	0.49
sentiment_nrc	0.72	0.28	0.40
syuzhet_nrc	0.71	0.27	0.39
sentiment_socal_google	0.70	0.18	0.28

Table 92 MMM 2016 Negative Precision and Recall Outcome (MR1)

Dictionaries	Precision	Recall	F1 Measure
sentiment_jockers_rinker	0.95	0.69	0.80
syuzhet_jockers	0.95	0.68	0.79
sentimentr_jockers	0.95	0.68	0.79
sentiment_sentistrength	0.96	0.66	0.78
sentiment_vadar	0.95	0.64	0.77
syuzhet_afinn	0.94	0.64	0.76
combined_dictionary	0.92	0.57	0.70
sentimentr_huliu	0.96	0.54	0.69
sentiment_berkeley	0.92	0.55	0.69
syuzhet_bing	0.96	0.50	0.66
sentiment_stanford	0.85	0.48	0.61
sentiment_senticnet	0.89	0.46	0.60
sentiment_nrc	0.97	0.43	0.59
sentiment_slangsd	0.88	0.44	0.59
sentimentr_sentiword	0.89	0.44	0.59
syuzhet_nrc	0.97	0.41	0.58
sentiment_inquirer	0.95	0.38	0.55
sentiment_loughran_mcdonald	0.94	0.38	0.54
sentiment_socal_google	0.94	0.24	0.39

Table 93 Dover 2016 Negative Precision and Recall Outcome (MR1)

Dictionaries	Precision	Recall	F1 Measure
syuzhet_jockers	0.81	0.57	0.67
sentiment_stanford	0.74	0.61	0.67
sentiment_sentistrength	0.81	0.57	0.67
sentimentr_jockers	0.81	0.57	0.67
sentiment_jockers_rinker	0.79	0.57	0.66
sentiment_vadar	0.81	0.45	0.58
syuzhet_afinn	0.79	0.45	0.58
sentimentr_huliu	0.83	0.43	0.57
syuzhet_bing	0.84	0.41	0.55
combined_dictionary	0.73	0.45	0.55
sentiment_berkeley	0.69	0.45	0.54
sentimentr_sentiword	0.64	0.46	0.53
sentiment_nrc	0.82	0.39	0.53
syuzhet_nrc	0.82	0.39	0.53
sentiment_senticnet	0.61	0.42	0.50
sentiment_slangsd	0.52	0.47	0.49
sentiment_loughran_mcdonald	0.78	0.34	0.48
sentiment_inquirer	0.76	0.32	0.45
sentiment_socal_google	0.77	0.16	0.27

Table 94 Anti-Austerity 2016 Negative Precision and Recall Outcome (MR1)

10.11.1.2 Neutral

Dictionaries	Precision	Recall	F1 Measure
sentiment_vadar	0.67	0.70	0.68
sentimentr_huliu	0.78	0.59	0.67
syuzhet_bing	0.79	0.57	0.66
syuzhet_afinn	0.69	0.64	0.66
sentiment_inquirer	0.78	0.56	0.65
sentimentr_jockers	0.53	0.82	0.65
sentiment_loughran_mcdonald	0.82	0.53	0.65
syuzhet_jockers	0.53	0.83	0.65
sentiment_jockers_rinker	0.52	0.84	0.64
sentiment_nrc	0.64	0.55	0.59
syuzhet_nrc	0.64	0.53	0.58
sentiment_socal_google	0.67	0.47	0.55
sentiment_slangsd	0.53	0.46	0.49
sentiment_stanford	0.57	0.42	0.49
combined_dictionary	0.45	0.52	0.48
sentiment_sentistrength	0.46	0.38	0.41
sentimentr_sentiword	0.22	0.65	0.33
sentiment_senticnet	0.05	0.54	0.09
sentiment_berkeley	0.04	0.29	0.07

Table 95 MMM 2015 Neutral Precision and Recall Outcome (MR1)

Dictionaries	Precision	Recall	F1 Measure
sentiment_sentistrength	0.74	0.65	0.69
sentiment_loughran_mcdonald	0.81	0.59	0.68
syuzhet_afinn	0.66	0.70	0.68
sentiment_vadar	0.62	0.75	0.68
sentimentr_huliu	0.73	0.63	0.67
syuzhet_bing	0.73	0.62	0.67
sentiment_inquirer	0.75	0.60	0.67
sentimentr_jockers	0.49	0.81	0.61
sentiment_jockers_rinker	0.48	0.83	0.60
syuzhet_jockers	0.48	0.80	0.60
sentiment_socal_google	0.64	0.54	0.59
sentiment_nrc	0.55	0.62	0.59
syuzhet_nrc	0.55	0.60	0.57
sentiment_stanford	0.57	0.52	0.54
sentiment_slangsd	0.52	0.53	0.52
combined_dictionary	0.40	0.55	0.47
sentimentr_sentiword	0.15	0.58	0.24
sentiment_berkeley	0.08	0.33	0.13
sentiment_senticnet	0.06	0.72	0.12

Table 96 MMM 2016 Neutral Precision and Recall Outcome (MR1)

Dictionaries	Precision	Recall	F1 Measure
sentiment_sentistrength	0.62	0.34	0.44
sentiment_vadar	0.46	0.40	0.43
sentimentr_jockers	0.38	0.46	0.41
syuzhet_jockers	0.36	0.46	0.41
syuzhet_afinn	0.51	0.33	0.40
sentiment_jockers_rinker	0.34	0.45	0.39
syuzhet_bing	0.61	0.27	0.38
sentimentr_huliu	0.56	0.28	0.38
sentiment_inquirer	0.66	0.25	0.37
sentiment_nrc	0.62	0.25	0.36
syuzhet_nrc	0.61	0.24	0.34
sentiment_loughran_mcdonald	0.80	0.22	0.34
sentiment_slangsd	0.60	0.22	0.32
sentiment_socal_google	0.59	0.21	0.31
combined_dictionary	0.41	0.25	0.31
sentiment_stanford	0.46	0.17	0.25
sentimentr_sentiword	0.18	0.29	0.22
sentiment_senticnet	0.07	0.36	0.11
sentiment_berkeley	0.08	0.17	0.10

Table 97 Dover 2016 Neutral Precision and Recall Outcome (MR1)

Dictionaries	Precision	Recall	F1 Measure
sentiment_sentistrength	0.74	0.64	0.69
sentimentr_huliu	0.70	0.62	0.66
sentiment_loughran_mcdonald	0.86	0.53	0.66
syuzhet_bing	0.71	0.60	0.65
sentiment_inquirer	0.71	0.59	0.65
sentiment_stanford	0.65	0.60	0.63
syuzhet_afinn	0.60	0.64	0.61
sentiment_vadar	0.55	0.68	0.61
sentiment_socal_google	0.63	0.52	0.57
sentiment_nrc	0.56	0.57	0.56
sentimentr_jockers	0.45	0.73	0.56
syuzhet_jockers	0.45	0.73	0.56
syuzhet_nrc	0.56	0.55	0.56
sentiment_jockers_rinker	0.44	0.74	0.55
sentiment_slangsd	0.51	0.49	0.50
combined_dictionary	0.40	0.57	0.47
sentimentr_sentiword	0.14	0.61	0.23
sentiment_senticnet	0.05	0.74	0.10
sentiment_berkeley	0.03	0.31	0.06

Table 98 Anti-Austerity 2016 Neutral Precision and Recall Outcome (MR1)

10.11.1.3 Positive

Dictionaries	Precision	Recall	F1 Measure
syuzhet_bing	0.67	0.27	0.39
sentimentr_huliu	0.65	0.26	0.38
syuzhet_afinn	0.70	0.24	0.36
sentiment_vadar	0.84	0.22	0.35
sentiment_inquirer	0.69	0.23	0.34
sentimentr_jockers	0.83	0.21	0.33
sentiment_jockers_rinker	0.81	0.21	0.33
syuzhet_jockers	0.81	0.20	0.33
sentiment_loughran_mcdonald	0.22	0.28	0.25
combined_dictionary	0.57	0.13	0.21
syuzhet_nrc	0.65	0.10	0.17
sentiment_nrc	0.65	0.10	0.17
sentiment_senticnet	0.81	0.09	0.16
sentimentr_sentiword	0.58	0.08	0.14
sentiment_berkeley	0.80	0.08	0.14
sentiment_sentistrength	0.17	0.08	0.11
sentiment_socal_google	0.33	0.06	0.10
sentiment_stanford	0.10	0.08	0.09
sentiment_slangsd	0.13	0.05	0.07

Table 99 MMM 2015 Positive Precision and Recall Outcome (MR1)

Dictionaries	Precision	Recall	F1 Measure
sentiment_sentistrength	0.65	0.40	0.50
syuzhet_bing	0.73	0.31	0.43
sentimentr_huliu	0.74	0.30	0.43
sentiment_inquirer	0.71	0.30	0.42
syuzhet_afinn	0.77	0.28	0.41
sentiment_vadar	0.87	0.25	0.39
sentiment_loughran_mcdonald	0.35	0.39	0.37
sentiment_jockers_rinker	0.86	0.23	0.37
sentimentr_jockers	0.85	0.23	0.36
syuzhet_jockers	0.83	0.23	0.36
combined_dictionary	0.62	0.18	0.27
sentiment_socal_google	0.65	0.16	0.26
sentiment_nrc	0.74	0.15	0.25
syuzhet_nrc	0.71	0.15	0.25
sentimentr_sentiword	0.73	0.13	0.22
sentiment_senticnet	0.80	0.12	0.21
sentiment_stanford	0.21	0.17	0.19
sentiment_berkeley	0.74	0.10	0.18
sentiment_slangsd	0.09	0.07	0.07

Table 100 MMM 2016 Positive Precision and Recall Outcome (MR1)

Dictionaries	Precision	Recall	F1 Measure
sentiment_sentistrength	0.79	0.11	0.19
syuzhet_bing	0.97	0.09	0.16
sentiment_loughran_mcdonald	0.34	0.09	0.15
combined_dictionary	0.86	0.07	0.13
sentimentr_huliu	0.83	0.07	0.13
syuzhet_afinn	0.69	0.07	0.12
sentiment_vadar	0.86	0.06	0.12
syuzhet_jockers	0.90	0.06	0.11
sentimentr_jockers	0.90	0.06	0.11
sentiment_stanford	0.38	0.06	0.11
sentiment_jockers_rinker	0.86	0.06	0.11
sentiment_inquirer	0.72	0.06	0.10
sentiment_nrc	0.55	0.04	0.08
syuzhet_nrc	0.55	0.04	0.08
sentiment_socal_google	0.69	0.04	0.07
sentiment_berkeley	0.83	0.04	0.07
sentiment_senticnet	0.90	0.03	0.06
sentimentr_sentiword	0.79	0.03	0.06
sentiment_slangsd	0.00	0.00	0.00

Table 101 Dover 2016 Positive Precision and Recall Outcome (MR1)

Dictionaries	Precision	Recall	F1 Measure
sentiment_sentistrength	0.78	0.42	0.55
sentimentr_huliu	0.83	0.28	0.42
syuzhet_bing	0.79	0.29	0.42
sentiment_stanford	0.52	0.34	0.41
syuzhet_afinn	0.88	0.25	0.39
sentiment_inquirer	0.78	0.25	0.38
sentiment_loughran_mcdonald	0.37	0.37	0.37
syuzhet_jockers	0.95	0.22	0.36
sentimentr_jockers	0.94	0.22	0.35
sentiment_jockers_rinker	0.93	0.22	0.35
sentiment_vadar	0.91	0.22	0.35
syuzhet_nrc	0.75	0.20	0.32
combined_dictionary	0.83	0.19	0.31
sentiment_nrc	0.73	0.20	0.31
sentiment_berkeley	0.91	0.13	0.23
sentiment_senticnet	0.89	0.13	0.22
sentimentr_sentiword	0.78	0.13	0.22
sentiment_socal_google	0.55	0.13	0.22
sentiment_slangsd	0.09	0.07	0.08

Table 102 Anti-Austerity 2016 Positive Precision and Recall Outcome (MR1)

10.11.2 MR2 Results

10.11.2.1 Negative

NEGATIVE	Precision	Recall	F-measure
syuzhet_jockers	0.24	0.61	0.35
sentiment_sentistrength	0.24	0.59	0.34
combined_dictionary	0.24	0.54	0.34
syuzhet_bing	0.26	0.47	0.34
sentimentr_jockers	0.23	0.58	0.33
sentiment_jockers_rinker	0.23	0.59	0.33
sentiment_stanford	0.22	0.65	0.33
sentimentr_huliu	0.25	0.47	0.33
syuzhet_nrc	0.26	0.44	0.32
sentiment_nrc_cat	0.25	0.44	0.32
sentiment_vadar	0.24	0.48	0.32
syuzhet_afinn	0.23	0.47	0.30
sentiment_berkeley	0.19	0.45	0.27
sentiment_senticnet	0.18	0.44	0.25
sentiment_inquirer	0.20	0.31	0.24
sentimentr_sentiword	0.16	0.42	0.23
sentiment_loughran_mcdonald	0.19	0.30	0.23
sentiment_slangsd	0.14	0.44	0.21
sentiment_socal_google	0.18	0.14	0.15

Table 103 2016 Anti-Austerity Negative Precision and Recall Outcome (MR2)

Category	Precision	Recall	F-measure
sentiment_jockers_rinker	0.79	0.72	0.76
sentimentr_jockers	0.79	0.71	0.75
syuzhet_jockers	0.78	0.71	0.74
sentiment_sentistrength	0.79	0.69	0.74
sentiment_vadar	0.79	0.67	0.73
syuzhet_afinn	0.78	0.67	0.72
combined_dictionary	0.76	0.59	0.66
sentimentr_huliu	0.79	0.56	0.66
sentiment_berkeley	0.74	0.56	0.64
syuzhet_bing	0.79	0.52	0.63
sentiment_nrc	0.80	0.45	0.58
sentiment_senticnet	0.72	0.47	0.57
syuzhet_nrc	0.80	0.43	0.56
sentiment_stanford	0.67	0.47	0.55
sentimentr_sentiword	0.72	0.45	0.55
sentiment_inquirer	0.79	0.40	0.54
sentiment_slangsd	0.67	0.43	0.52
sentiment_loughran_mcdonald	0.74	0.38	0.50
sentiment_socal_google	0.77	0.25	0.38

Table 104 2016 Dover Negative Precision and Recall Outcome (MR2)

Dictionary	Precisio	Recal	F-measur
sentiment_jockers_rinker	0.46	0.71	0.56
sentimentr_jockers	0.46	0.69	0.56
syuzhet_jockers	0.46	0.70	0.55
sentiment_vadar	0.50	0.60	0.55
syuzhet_afinn	0.49	0.60	0.54
combined_dictionary	0.44	0.59	0.50
sentiment_sentistrength	0.44	0.55	0.49
syuzhet_bing	0.47	0.44	0.46
sentiment_loughran_mcdonal	0.44	0.46	0.45
sentimentr_huliu	0.46	0.44	0.45
sentiment_senticnet	0.37	0.54	0.44
sentiment_inquirer	0.45	0.37	0.40
sentiment_berkeley	0.41	0.40	0.40
sentimentr_sentiword	0.33	0.48	0.39
sentiment_stanford	0.32	0.45	0.37
sentiment_slangsd	0.30	0.48	0.37
sentiment_nrc	0.42	0.27	0.33
syuzhet_nrc	0.42	0.26	0.32
sentiment_socal_google	0.38	0.16	0.23

Table 105 2016 MMM Negative Precision and Recall Outcome (MR2)

Dictionary	Precision	Recall	F-Measure
syuzhet_jockers	0.55	0.81	0.65
sentiment_jockers_rinker	0.54	0.81	0.65
sentimentr_jockers	0.55	0.80	0.65
syuzhet_afinn	0.57	0.67	0.62
sentiment_vadar	0.57	0.66	0.61
sentimentr_huliu	0.58	0.56	0.57
combined_dictionary	0.51	0.61	0.56
syuzhet_bing	0.58	0.53	0.55
sentiment_senticnet	0.48	0.62	0.54
sentiment_berkeley	0.54	0.53	0.53
sentimentr_sentiword	0.47	0.60	0.53
sentiment_inquirer	0.58	0.48	0.52
sentiment_loughran_mcdonald	0.51	0.52	0.52
sentiment_stanford	0.39	0.45	0.42
sentiment_sentistrength	0.38	0.44	0.41
sentiment_slangsd	0.37	0.41	0.39
sentiment_nrc_cat	0.54	0.26	0.35
syuzhet_nrc	0.54	0.25	0.34
sentiment_socal_google	0.51	0.21	0.30

Table 106 2015 MMM Negative Precision and Recall Outcome (MR2)

10.11.2.2 Neutral

NEUTRAL	Precision	Recall	F-measure
sentiment_loughran_mcdonald	0.73	0.80	0.77
syuzhet_bing	0.57	0.85	0.68
sentiment_inquirer	0.56	0.83	0.67
sentimentr_huliu	0.55	0.86	0.67
sentiment_socal_google	0.56	0.82	0.66
sentiment_sentistrength	0.55	0.84	0.66
sentiment_stanford	0.51	0.84	0.63
syuzhet_nrc	0.47	0.83	0.60
sentiment_nrc_cat	0.47	0.83	0.60
syuzhet_afinn	0.45	0.86	0.59
sentiment_slangsd	0.46	0.77	0.58
sentiment_vadar	0.40	0.88	0.55
combined_dictionary	0.34	0.85	0.49
syuzhet_jockers	0.32	0.91	0.47
sentimentr_jockers	0.32	0.91	0.47
sentiment_jockers_rinker	0.31	0.91	0.46
sentimentr_sentiword	0.11	0.86	0.20
sentiment_berkeley	0.04	0.72	0.08
sentiment_senticnet	0.04	0.93	0.07

Table 107 2016 Anti-Austerity Neutral Precision and Recall Outcome (MR2)

Dictionary	Precision	Recall	F-measure
sentiment_loughran_mcdonald	0.69	0.37	0.48
sentiment_inquirer	0.55	0.42	0.48
sentiment_nrc	0.51	0.41	0.46
sentiment_sentistrength	0.43	0.48	0.45
syuzhet_nrc	0.51	0.40	0.45
syuzhet_bing	0.45	0.40	0.43
syuzhet_afinn	0.38	0.48	0.42
sentimentr_huliu	0.41	0.42	0.42
sentiment_slangsd	0.49	0.35	0.41
sentiment_socal_google	0.47	0.34	0.40
sentiment_vadar	0.31	0.53	0.39
sentiment_stanford	0.41	0.31	0.35
combined_dictionary	0.31	0.37	0.34
sentimentr_jockers	0.23	0.55	0.32
syuzhet_jockers	0.21	0.55	0.31
sentiment_jockers_rinker	0.21	0.56	0.31
sentimentr_sentiword	0.12	0.39	0.18
sentiment_berkeley	0.07	0.29	0.11
sentiment_senticnet	0.04	0.42	0.07

Table 108 2016 Dover Neutral Precision and Recall Outcome (MR2)

Dictionary	Precision	Recall	F-measure
sentiment_loughran_mcdonald	0.73	0.70	0.71
sentiment_inquirer	0.68	0.72	0.70
sentimentr_huliu	0.64	0.74	0.69
syuzhet_bing	0.65	0.73	0.68
sentiment_sentistrength	0.62	0.72	0.67
syuzhet_afinn	0.56	0.79	0.66
sentiment_socal_google	0.61	0.69	0.65
sentiment_vadar	0.52	0.83	0.64
sentiment_stanford	0.54	0.64	0.59
syuzhet_nrc	0.50	0.71	0.58
sentiment_nrc	0.49	0.72	0.58
sentiment_slangsd	0.49	0.67	0.56
sentimentr_jockers	0.40	0.88	0.55
syuzhet_jockers	0.39	0.87	0.54
sentiment_jockers_rinker	0.39	0.89	0.54
combined_dictionary	0.37	0.68	0.48
sentimentr_sentiword	0.13	0.67	0.22
sentiment_berkeley	0.10	0.51	0.17
sentiment_senticnet	0.05	0.80	0.10

Table 109 2016 MMM Neutral Precision and Recall Outcome (MR2)

Dictionary	Precision	Recall	F-Measure
syuzhet_bing	0.66	0.72	0.69
sentimentr_huliu	0.65	0.73	0.69
sentiment_inquirer	0.65	0.71	0.68
sentiment_loughran_mcdonald	0.68	0.66	0.67
syuzhet_afinn	0.56	0.78	0.66
sentiment_vadar	0.52	0.81	0.64
sentiment_socal_google	0.60	0.63	0.62
sentiment_nrc_cat	0.52	0.68	0.59
syuzhet_nrc	0.53	0.66	0.59
sentiment_slangsd	0.50	0.64	0.56
sentiment_stanford	0.53	0.58	0.55
syuzhet_jockers	0.39	0.90	0.54
sentimentr_jockers	0.39	0.89	0.54
sentiment_sentistrength	0.48	0.58	0.53
sentiment_jockers_rinker	0.37	0.90	0.52
combined_dictionary	0.38	0.67	0.49
sentimentr_sentiword	0.18	0.77	0.29
sentiment_berkeley	0.05	0.52	0.08
sentiment_senticnet	0.04	0.71	0.08

Table 110 2015 MMM Neutral Precision and Recall Outcome (MR2)

10.11.2.3 Positive

Positive	Precision	Recall	F-measure
sentiment_sentistrength	0.49	0.30	0.37
syuzhet_afinn	0.73	0.24	0.36
sentimentr_huliu	0.63	0.24	0.35
syuzhet_bing	0.60	0.24	0.35
sentiment_stanford	0.40	0.30	0.34
sentiment_inquirer	0.60	0.22	0.32
sentimentr_jockers	0.77	0.20	0.32
sentiment_vadar	0.74	0.20	0.32
sentiment_jockers_rinker	0.76	0.20	0.32
syuzhet_jockers	0.76	0.20	0.31
syuzhet_nrc	0.63	0.19	0.29
sentiment_nrc	0.63	0.19	0.29
sentiment_loughran_mcdonald	0.27	0.30	0.29
combined_dictionary	0.69	0.18	0.28
sentiment_berkeley	0.83	0.13	0.23
sentiment_senticnet	0.81	0.13	0.23
sentimentr_sentiword	0.69	0.13	0.22
sentiment_socal_google	0.45	0.12	0.19
sentiment_slangsd	0.06	0.06	0.06

Table 111 2016 Anti-Austerity Positive Precision and Recall Outcome (MR2)

Dictionary	Precision	Recall	F-measure
sentiment_sentistrength	0.53	0.12	0.20
sentiment_loughran_mcdonal	0.29	0.13	0.18
syuzhet_afinn	0.61	0.10	0.17
sentimentr_jockers	0.78	0.09	0.16
sentiment_jockers_rinker	0.76	0.09	0.16
sentiment_vadar	0.69	0.09	0.15
combined_dictionary	0.61	0.09	0.15
syuzhet_jockers	0.73	0.08	0.15
sentimentr_huliu	0.59	0.09	0.15
syuzhet_bing	0.55	0.08	0.14
sentiment_inquirer	0.57	0.07	0.13
sentiment_nrc	0.53	0.07	0.13
sentiment_stanford	0.27	0.08	0.12
syuzhet_nrc	0.49	0.07	0.12
sentiment_berkeley	0.78	0.06	0.11
sentiment_senticnet	0.80	0.05	0.09
sentimentr_sentiword	0.67	0.04	0.08
sentiment_socal_google	0.37	0.03	0.06
sentiment_slangsd	0.06	0.01	0.02

Table 112 2016 Dover Positive Precision and Recall Outcome (MR2)

Dictionary	Precision	Recall	F-measure
sentiment_jockers_rinker	0.72	0.28	0.40
sentimentr_jockers	0.72	0.27	0.39
sentimentr_huliu	0.61	0.35	0.44
sentimentr_sentiword	0.69	0.17	0.27
sentiment_nrc	0.60	0.17	0.27
sentiment_loughran_mcdonald	0.24	0.37	0.29
sentiment_senticnet	0.78	0.16	0.27
sentiment_inquirer	0.54	0.32	0.40
sentiment_slangsd	0.11	0.12	0.12
sentiment_socal_google	0.53	0.19	0.28
sentiment_vadar	0.72	0.30	0.42
sentiment_stanford	0.16	0.18	0.17
syuzhet_jockers	0.70	0.27	0.39
syuzhet_bing	0.58	0.34	0.43
syuzhet_afinn	0.58	0.30	0.39
syuzhet_nrc	0.60	0.18	0.27
sentiment_berkeley	0.69	0.13	0.22
sentiment_sentistrength	0.43	0.38	0.40
combined_dictionary	0.53	0.21	0.30

Table 113 2016 MMM Positive Precision and Recall Outcome (MR2)

Dictionary	Precision	Recall	F-Measure
sentiment_jockers_rinker	0.60	0.17	0.27
sentimentr_jockers	0.61	0.17	0.27
sentimentr_huliu	0.47	0.22	0.30
sentimentr_sentiword	0.48	0.08	0.13
sentiment_nrc_cat	0.57	0.10	0.17
sentiment_loughran_mcdonald	0.13	0.19	0.16
sentiment_senticnet	0.69	0.09	0.16
sentiment_inquirer	0.47	0.18	0.26
sentiment_slangsd	0.16	0.07	0.10
sentiment_socal_google	0.32	0.07	0.12
sentiment_vadar	0.65	0.20	0.30
sentiment_stanford	0.10	0.09	0.10
syuzhet_jockers	0.62	0.18	0.28
syuzhet_bing	0.48	0.22	0.31
syuzhet_afinn	0.49	0.20	0.28
syuzhet_nrc	0.58	0.10	0.17
sentiment_berkeley	0.70	0.08	0.14
sentiment_sentistrength	0.14	0.08	0.10
combined_dictionary	0.42	0.11	0.18

Table 114 2015 MMM Positive Precision and Recall Outcome (MR2)

10.11.3 MR3 Results

10.11.3.1 Negative

Dictionaries	Precision	Recall	F-Measure
sentiment_jockers_rinker	0.07	0.62	0.13
sentimentr_jockers	0.07	0.62	0.13
sentimentr_huliu	0.07	0.41	0.12
sentimentr_sentiword	0.06	0.44	0.10
sentiment_nrc	0.08	0.22	0.11
sentiment_loughran_mcdonald	0.04	0.26	0.07
sentiment_senticnet	0.07	0.52	0.12
sentiment_inquirer	0.08	0.39	0.13
sentiment_slangsd	0.05	0.34	0.09
sentiment_socal_google	0.07	0.18	0.10
sentiment_vadar	0.07	0.49	0.12
sentiment_stanford	0.06	0.41	0.11
syuzhet_jockers	0.07	0.62	0.13
syuzhet_bing	0.08	0.41	0.13
syuzhet_afinn	0.08	0.56	0.14
syuzhet_nrc	0.08	0.22	0.12
sentiment_berkeley	0.07	0.40	0.12
sentiment_sentistrength	0.06	0.43	0.11
combined_dictionary	0.06	0.43	0.11

Table 115 2015 MMM Negative Precision and Recall Outcome (MR3)

Dictionaries	Precision	Recall	F-Measure
sentiment_jockers_rinker	0.03	0.73	0.07
sentimentr_jockers	0.04	0.77	0.07
sentimentr_huliu	0.05	0.67	0.10
sentimentr_sentiword	0.03	0.50	0.05
sentiment_nrc	0.05	0.47	0.10
sentiment_loughran_mcdonald	0.02	0.27	0.04
sentiment_senticnet	0.03	0.63	0.06
sentiment_inquirer	0.05	0.50	0.08
sentiment_slangsd	0.02	0.40	0.03
sentiment_socal_google	0.05	0.30	0.09
sentiment_vadar	0.05	0.83	0.10
sentiment_stanford	0.02	0.47	0.05
syuzhet_jockers	0.04	0.77	0.07
syuzhet_bing	0.06	0.70	0.10
syuzhet_afinn	0.04	0.70	0.08
syuzhet_nrc	0.06	0.47	0.10
sentiment_berkeley	0.05	0.63	0.09
sentiment_sentistrength	0.05	0.90	0.10
combined_dictionary	0.04	0.70	0.07

Table 116 2016 MMM Negative Precision and Recall Outcome (MR3)

Dictionaries	Precision	Recall	F-Measure
sentiment_jockers_rinker	0.07	0.71	0.13
sentimentr_jockers	0.08	0.74	0.14
sentimentr_huliu	0.08	0.63	0.15
sentimentr_sentiword	0.08	0.53	0.14
sentiment_nrc	0.08	0.47	0.14
sentiment_loughran_mcdonald	0.04	0.23	0.07
sentiment_senticnet	0.07	0.46	0.12
sentiment_inquirer	0.07	0.40	0.12
sentiment_slangsd	0.07	0.46	0.12
sentiment_socal_google	0.09	0.31	0.14
sentiment_vadar	0.08	0.71	0.14
sentiment_stanford	0.08	0.57	0.13
syuzhet_jockers	0.07	0.73	0.13
syuzhet_bing	0.09	0.63	0.16
syuzhet_afinn	0.08	0.73	0.14
syuzhet_nrc	0.08	0.48	0.14
sentiment_berkeley	0.07	0.53	0.12
sentiment_sentistrength	0.08	0.77	0.15
combined_dictionary	0.07	0.60	0.13

Table 117 2016 Dover Negative Precision and Recall Outcome (MR3)

Dictionaries	Precision	Recall	F-Measure
sentiment_jockers_rinker	0.05	0.59	0.09
sentimentr_jockers	0.05	0.59	0.09
sentimentr_huliu	0.06	0.54	0.11
sentimentr_sentiword	0.04	0.46	0.07
sentiment_nrc	0.04	0.32	0.07
sentiment_loughran_mcdonald	0.02	0.15	0.03
sentiment_senticnet	0.04	0.54	0.08
sentiment_inquirer	0.02	0.15	0.04
sentiment_slangsd	0.02	0.39	0.05
sentiment_socal_google	0.03	0.10	0.04
sentiment_vadar	0.05	0.51	0.10
sentiment_stanford	0.05	0.73	0.09
syuzhet_jockers	0.05	0.61	0.09
syuzhet_bing	0.06	0.54	0.11
syuzhet_afinn	0.06	0.63	0.12
syuzhet_nrc	0.03	0.27	0.06
sentiment_berkeley	0.03	0.32	0.05
sentiment_sentistrength	0.04	0.54	0.08
combined_dictionary	0.05	0.54	0.09

Table 118 2016 Anti-Austerity Negative Precision and Recall Outcome (MR3)

10.11.3.2 Neutral

Dictionaries	Precision	Recall	F-Measure
sentiment_jockers_rinker	0.25	0.96	0.40
sentimentr_jockers	0.26	0.96	0.41
sentimentr_huliu	0.52	0.93	0.67
sentimentr_sentiword	0.14	0.97	0.25
sentiment_nrc	0.45	0.93	0.61
sentiment_loughran_mcdonald	0.60	0.91	0.72
sentiment_senticnet	0.04	1.00	0.07
sentiment_inquirer	0.55	0.95	0.70
sentiment_slangsd	0.45	0.91	0.60
sentiment_socal_google	0.55	0.92	0.69
sentiment_vadar	0.39	0.96	0.55
sentiment_stanford	0.52	0.91	0.66
syuzhet_jockers	0.26	0.96	0.41
syuzhet_bing	0.55	0.93	0.69
syuzhet_afinn	0.44	0.95	0.60
syuzhet_nrc	0.47	0.93	0.62
sentiment_berkeley	0.05	0.88	0.09
sentiment_sentistrength	0.47	0.91	0.62
combined_dictionary	0.33	0.93	0.49

Table 119 2015 MMM Neutral Precision and Recall Outcome (MR3)

Dictionaries	Precision	Recall	F-Measure
sentiment_jockers_rinker	0.27	1.00	0.43
sentimentr_jockers	0.29	1.00	0.45
sentimentr_huliu	0.54	0.99	0.70
sentimentr_sentiword	0.12	0.98	0.22
sentiment_nrc	0.42	0.99	0.59
sentiment_loughran_mcdonald	0.64	0.97	0.77
sentiment_senticnet	0.04	1.00	0.08
sentiment_inquirer	0.59	0.99	0.74
sentiment_slangsd	0.45	0.98	0.62
sentiment_socal_google	0.55	0.98	0.71
sentiment_vadar	0.40	1.00	0.57
sentiment_stanford	0.51	0.98	0.67
syuzhet_jockers	0.28	1.00	0.44
syuzhet_bing	0.55	0.99	0.71
syuzhet_afinn	0.45	0.99	0.62
syuzhet_nrc	0.44	0.99	0.61
sentiment_berkeley	0.12	0.97	0.21
sentiment_sentistrength	0.54	0.99	0.70
combined_dictionary	0.34	0.99	0.51

Table 120 2016 MMM Neutral Precision and Recall Outcome (MR3)

Dictionaries	Precision	Recall	F-Measure
sentiment_jockers_rinker	0.12	0.94	0.22
sentimentr_jockers	0.13	0.95	0.23
sentimentr_huliu	0.32	0.95	0.48
sentimentr_sentiword	0.10	0.95	0.18
sentiment_nrc	0.40	0.94	0.56
sentiment_loughran_mcdonald	0.59	0.92	0.72
sentiment_senticnet	0.03	0.96	0.06
sentiment_inquirer	0.42	0.94	0.58
sentiment_slangsd	0.45	0.93	0.60
sentiment_socal_google	0.45	0.94	0.61
sentiment_vadar	0.19	0.96	0.32
sentiment_stanford	0.43	0.94	0.59
syuzhet_jockers	0.13	0.95	0.22
syuzhet_bing	0.36	0.95	0.53
syuzhet_afinn	0.26	0.95	0.40
syuzhet_nrc	0.41	0.94	0.57
sentiment_berkeley	0.07	0.92	0.13
sentiment_sentistrength	0.30	0.97	0.46
combined_dictionary	0.27	0.93	0.42

Table 121 2016 Dover Neutral Precision and Recall Outcome (MR3)

Dictionaries	Precision	Recall	F-Measure
sentiment_jockers_rinker	0.27	0.97	0.42
sentimentr_jockers	0.28	0.96	0.43
sentimentr_huliu	0.50	0.97	0.66
sentimentr_sentiword	0.11	0.97	0.19
sentiment_nrc	0.43	0.94	0.59
sentiment_loughran_mcdonald	0.70	0.94	0.80
sentiment_senticnet	0.03	1.00	0.06
sentiment_inquirer	0.52	0.95	0.67
sentiment_slangsd	0.45	0.93	0.61
sentiment_socal_google	0.52	0.93	0.67
sentiment_vadar	0.36	0.96	0.52
sentiment_stanford	0.48	0.95	0.63
syuzhet_jockers	0.28	0.97	0.43
syuzhet_bing	0.53	0.96	0.68
syuzhet_afinn	0.42	0.97	0.58
syuzhet_nrc	0.44	0.94	0.60
sentiment_berkeley	0.05	0.93	0.09
sentiment_sentistrength	0.51	0.95	0.66
combined_dictionary	0.31	0.95	0.47

Table 122 2016 Anti-Austerity Neutral Precision and Recall Outcome (MR3)

10.11.3.3 Positive

Dictionaries	Precision	Recall	F-Measure
sentiment_jockers_rinker	0.47	0.06	0.10
sentimentr_jockers	0.47	0.06	0.10
sentimentr_huliu	0.28	0.06	0.09
sentimentr_sentiword	0.49	0.03	0.06
sentiment_nrc	0.53	0.04	0.07
sentiment_loughran_mcdonald	0.16	0.10	0.13
sentiment_senticnet	0.58	0.03	0.06
sentiment_inquirer	0.40	0.07	0.11
sentiment_slangsd	0.14	0.03	0.04
sentiment_socal_google	0.23	0.02	0.04
sentiment_vadar	0.49	0.06	0.11
sentiment_stanford	0.12	0.05	0.07
syuzhet_jockers	0.47	0.06	0.10
syuzhet_bing	0.33	0.07	0.11
syuzhet_afinn	0.35	0.06	0.10
syuzhet_nrc	0.49	0.04	0.07
sentiment_berkeley	0.63	0.03	0.06
sentiment_sentistrength	0.09	0.02	0.04
combined_dictionary	0.30	0.03	0.06

Table 123 2015 MMM Positive Precision and Recall Outcome (MR3)

Dictionaries	Precision	Recall	F-Measure
sentiment_jockers_rinker	0.92	0.02	0.05
sentimentr_jockers	0.92	0.02	0.05
sentimentr_huliu	0.83	0.03	0.06
sentimentr_sentiword	0.67	0.01	0.02
sentiment_nrc	0.92	0.02	0.04
sentiment_loughran_mcdonald	0.17	0.02	0.03
sentiment_senticnet	0.92	0.01	0.03
sentiment_inquirer	0.58	0.02	0.04
sentiment_slangsd	0.08	0.01	0.01
sentiment_socal_google	0.50	0.01	0.02
sentiment_vadar	0.92	0.03	0.05
sentiment_stanford	0.25	0.02	0.04
syuzhet_jockers	0.83	0.02	0.04
syuzhet_bing	0.58	0.02	0.04
syuzhet_afinn	0.67	0.02	0.04
syuzhet_nrc	0.83	0.02	0.03
sentiment_berkeley	0.67	0.01	0.02
sentiment_sentistrength	0.67	0.04	0.07
combined_dictionary	0.67	0.02	0.03

Table 124 2016 MMM Positive Precision and Recall Outcome (MR3)

Dictionaries	Precision	Recall	F-Measure
sentiment_jockers_rinker	0.74	0.03	0.06
sentimentr_jockers	0.79	0.03	0.07
sentimentr_huliu	0.63	0.04	0.07
sentimentr_sentiword	0.63	0.02	0.03
sentiment_nrc	0.37	0.02	0.04
sentiment_loughran_mcdonald	0.21	0.04	0.06
sentiment_senticnet	0.95	0.02	0.04
sentiment_inquirer	0.47	0.02	0.05
sentiment_slangsd	0.05	0.00	0.01
sentiment_socal_google	0.74	0.03	0.05
sentiment_vadar	0.63	0.03	0.06
sentiment_stanford	0.37	0.04	0.07
syuzhet_jockers	0.74	0.03	0.06
syuzhet_bing	0.68	0.04	0.08
syuzhet_afinn	0.47	0.03	0.06
syuzhet_nrc	0.37	0.02	0.04
sentiment_berkeley	0.63	0.02	0.04
sentiment_sentistrength	0.42	0.04	0.07
combined_dictionary	0.63	0.03	0.07

Table 125 2016 Dover Positive Precision and Recall Outcome (MR3)

Dictionaries	Precision	Recall	F-Measure
sentiment_jockers_rinker	0.89	0.09	0.16
sentimentr_jockers	0.90	0.09	0.17
sentimentr_huliu	0.80	0.12	0.21
sentimentr_sentiword	0.77	0.06	0.11
sentiment_nrc	0.64	0.08	0.14
sentiment_loughran_mcdonald	0.33	0.14	0.20
sentiment_senticnet	0.85	0.05	0.10
sentiment_inquirer	0.72	0.10	0.18
sentiment_slangsd	0.05	0.02	0.03
sentiment_socal_google	0.49	0.05	0.10
sentiment_vadar	0.87	0.09	0.17
sentiment_stanford	0.51	0.15	0.23
syuzhet_jockers	0.90	0.09	0.17
syuzhet_bing	0.74	0.12	0.20
syuzhet_afinn	0.80	0.10	0.18
syuzhet_nrc	0.66	0.08	0.14
sentiment_berkeley	0.87	0.06	0.10
sentiment_sentistrength	0.64	0.15	0.24
combined_dictionary	0.79	0.08	0.15

Table 126 2016 Anti-Austerity Positive Precision and Recall Outcome (MR3)

10.11.4 Macro and Micro Precision and Recall

Dictionaries	Micro-Precisio	Micro-Reca	Micro-F-Score	Macro-Precisio	Macro-Recall	Macro-F-Score
sentiment_jockers_rinker	0.70	0.67	0.68	0.72	0.61	0.66
sentimentr_jockers	0.71	0.66	0.68	0.73	0.60	0.66
syuzhet_jockers	0.70	0.66	0.68	0.72	0.60	0.66
sentiment_vadar	0.76	0.57	0.65	0.78	0.51	0.62
syuzhet_afinn	0.76	0.57	0.65	0.74	0.50	0.60
sentimentr_huliu	0.81	0.53	0.64	0.77	0.46	0.58
syuzhet_bing	0.81	0.51	0.62	0.78	0.45	0.57
sentiment_inquirer	0.81	0.47	0.59	0.78	0.41	0.54
sentiment_loughran_mcdonald	0.77	0.52	0.62	0.61	0.44	0.52
combined dictionary	0.61	0.47	0.53	0.60	0.41	0.49
sentimentr_sentiword	0.49	0.40	0.44	0.50	0.43	0.47
sentiment_senticnet	0.44	0.36	0.39	0.53	0.41	0.46
sentiment_nrc	0.70	0.31	0.43	0.71	0.31	0.43
syuzhet_nrc	0.70	0.31	0.43	0.72	0.30	0.42
sentiment_berkeley	0.43	0.28	0.34	0.55	0.29	0.38
sentiment_stanford	0.56	0.41	0.48	0.43	0.32	0.37
sentiment_slangsd	0.53	0.39	0.45	0.42	0.31	0.36
sentiment_socal_google	0.66	0.28	0.39	0.58	0.24	0.34
sentiment_sentistrength	0.49	0.36	0.41	0.40	0.29	0.33

10.11.4.1 MR1 Results

Table 127 2015 MMM Micro/Macro/F-measure Precision and Recall (MR1)

Dictionaries	Micro-Precisio	Micro-Reca	Micro-F-Score	Macro-Precisio	Macro-Recall	Macro-F-Score
sentiment_jockers_rinker	0.62	0.58	0.60	0.68	0.58	0.63
sentimentr_jockers	0.62	0.57	0.59	0.69	0.56	0.62
sentiment_vadar	0.71	0.55	0.62	0.76	0.52	0.62
syuzhet_jockers	0.62	0.58	0.60	0.68	0.57	0.62
sentiment_sentistrength	0.73	0.58	0.64	0.71	0.53	0.61
syuzhet_afinn	0.71	0.56	0.63	0.73	0.51	0.61
sentimentr_huliu	0.74	0.50	0.60	0.74	0.46	0.56
syuzhet_bing	0.75	0.50	0.60	0.75	0.45	0.56
sentiment_loughran_mcdonald	0.73	0.52	0.61	0.63	0.48	0.54
sentiment_inquirer	0.75	0.46	0.57	0.74	0.42	0.54
sentiment_combined	0.54	0.45	0.49	0.57	0.43	0.49
sentiment_senticnet	0.36	0.31	0.33	0.49	0.45	0.47
sentiment_nrc	0.62	0.35	0.44	0.67	0.35	0.46
syuzhet_nrc	0.61	0.34	0.44	0.66	0.34	0.45
sentimentr_sentiword	0.37	0.33	0.35	0.48	0.39	0.43
sentiment_socal_google	0.65	0.32	0.43	0.66	0.29	0.41
sentiment_stanford	0.52	0.45	0.49	0.44	0.38	0.41
sentiment_berkeley	0.36	0.24	0.29	0.51	0.28	0.36
sentiment_slangsd	0.47	0.46	0.46	0.37	0.36	0.36

Table 128 2016 MMM Micro/Macro/F-measure Precision and Recall (MR1)

Dictionaries	Micro-Precisio	Micro-Reca	Micro-F-Score	Macro-Precisio	Macro-Recall	Macro-F-Score
syuzhet_jockers	0.83	0.52	0.64	0.73	0.40	0.52
sentimentr_jockers	0.83	0.51	0.63	0.74	0.40	0.52
sentiment_jockers_rinker	0.82	0.52	0.64	0.72	0.40	0.51
sentiment_sentistrength	0.88	0.53	0.66	0.79	0.37	0.50
sentiment_vadar	0.84	0.49	0.62	0.76	0.37	0.50
syuzhet_afinn	0.84	0.49	0.62	0.72	0.34	0.46
sentimentr_huliu	0.85	0.40	0.55	0.78	0.30	0.43
syuzhet_bing	0.87	0.38	0.53	0.85	0.29	0.43
sentiment_combined	0.80	0.41	0.55	0.73	0.29	0.42
sentiment_senticnet	0.67	0.29	0.40	0.62	0.28	0.39
sentimentr_sentiword	0.69	0.29	0.40	0.62	0.25	0.36
sentiment_nrc	0.85	0.32	0.46	0.71	0.24	0.36
sentiment_berkeley	0.72	0.36	0.48	0.61	0.25	0.36
sentiment_inquirer	0.85	0.29	0.43	0.78	0.23	0.36
syuzhet_nrc	0.84	0.30	0.45	0.71	0.23	0.35
sentiment_loughran_mcdonald	0.88	0.30	0.45	0.70	0.23	0.35
sentiment_stanford	0.74	0.35	0.47	0.57	0.24	0.34
sentiment_slangsd	0.78	0.33	0.46	0.49	0.22	0.30
sentiment_socal_google	0.78	0.19	0.31	0.74	0.16	0.27

Table 129 2016 Dover Micro/Macro/F-measure Precision and Recall (MR1)

Dictionaries	Micro-Precisio	Micro-Reca	Micro-F-Score	Macro-Precisio	Macro-Recall	Macro-F-Score
sentiment_sentistrength	0.77	0.58	0.66	0.78	0.54	0.64
syuzhet_jockers	0.65	0.48	0.55	0.74	0.51	0.60
sentimentr_jockers	0.64	0.48	0.55	0.73	0.51	0.60
sentiment_jockers_rinker	0.63	0.48	0.55	0.72	0.51	0.60
sentiment_stanford	0.68	0.57	0.62	0.64	0.52	0.57
sentimentr_huliu	0.76	0.47	0.58	0.78	0.45	0.57
sentiment_vadar	0.68	0.44	0.54	0.76	0.45	0.56
syuzhet_afinn	0.70	0.46	0.55	0.76	0.45	0.56
syuzhet_bing	0.76	0.46	0.58	0.78	0.43	0.56
sentiment_loughran_mcdonald	0.78	0.45	0.57	0.67	0.42	0.51
sentiment_inquirer	0.73	0.41	0.53	0.75	0.39	0.51
combined_dictionary	0.57	0.39	0.46	0.65	0.40	0.50
sentiment_nrc	0.66	0.40	0.50	0.70	0.39	0.50
syuzhet_nrc	0.66	0.39	0.49	0.71	0.38	0.50
sentiment_senticnet	0.36	0.27	0.31	0.52	0.43	0.47
sentimentr_sentiword	0.41	0.31	0.35	0.52	0.40	0.45
sentiment_socal_google	0.64	0.29	0.40	0.65	0.27	0.38
sentiment_berkeley	0.37	0.27	0.31	0.54	0.30	0.38
sentiment_slangsd	0.48	0.43	0.45	0.37	0.34	0.36

Table 130 2016 AA Micro/Macro/F-measure Precision and Recall (MR1)

10.11.4.2 MR2 Results

Dictionary	Micro-Precision	Micro-Rec	Micro-F-Score	Macro-Precisic	Macro-Recal	Macro-F-Score
sentimentr_huliu	0.49	0.61	0.54	0.47	0.52	0.50
syuzhet_bing	0.51	0.62	0.56	0.48	0.52	0.50
syuzhet_jockers	0.34	0.50	0.40	0.44	0.57	0.50
sentimentr_jockers	0.33	0.50	0.40	0.44	0.56	0.49
syuzhet_afinn	0.42	0.56	0.48	0.47	0.52	0.49
sentiment_jockers_rinker	0.32	0.50	0.39	0.43	0.57	0.49
sentiment_sentistrength	0.46	0.68	0.55	0.43	0.58	0.49
sentiment_vadar	0.39	0.51	0.45	0.46	0.52	0.49
syuzhet_nrc	0.44	0.53	0.48	0.45	0.49	0.47
combined_dictionary	0.35	0.48	0.40	0.42	0.52	0.47
sentiment_nrc_cat	0.44	0.53	0.48	0.45	0.49	0.47
sentiment_stanford	0.41	0.70	0.52	0.38	0.59	0.46
sentiment_inquirer	0.50	0.57	0.53	0.46	0.45	0.45
sentiment_loughran_mcdonald	0.58	0.68	0.63	0.40	0.47	0.43
sentiment_senticnet	0.14	0.21	0.17	0.34	0.50	0.41
sentiment_berkeley	0.15	0.22	0.18	0.36	0.43	0.39
sentimentr_sentiword	0.18	0.27	0.21	0.32	0.47	0.38
sentiment_socal_google	0.51	0.48	0.49	0.40	0.36	0.38
sentiment_slangsd	0.32	0.59	0.42	0.22	0.42	0.29

Table 131 2016 AA Micro/Macro/F-measure Precision and Recall (MR2)

Dictionary	Micro-Precisic	Micro-Reca	Micro-F-Sco	Macro-Precisi	Macro-Reca	Macro-F-Sco
sentiment_jockers_rinker	0.60	0.53	0.56	0.59	0.46	0.51
sentimentr_jockers	0.60	0.52	0.56	0.60	0.45	0.51
syuzhet_jockers	0.59	0.52	0.55	0.58	0.45	0.50
sentiment_vadar	0.62	0.51	0.56	0.60	0.43	0.50
sentiment_sentistrength	0.66	0.56	0.60	0.58	0.43	0.49
syuzhet_afinn	0.64	0.52	0.57	0.59	0.42	0.49
sentimentr_huliu	0.63	0.43	0.52	0.60	0.36	0.45
syuzhet_bing	0.64	0.41	0.50	0.60	0.34	0.43
combined_dictionary	0.59	0.44	0.50	0.56	0.35	0.43
sentiment_nrc	0.66	0.37	0.47	0.62	0.31	0.41
sentiment_inquirer	0.67	0.35	0.46	0.64	0.30	0.41
syuzhet_nrc	0.66	0.35	0.46	0.60	0.30	0.40
sentiment_senticnet	0.44	0.28	0.34	0.52	0.31	0.39
sentiment_loughran_mcdonald	0.70	0.36	0.48	0.57	0.29	0.39
sentiment_berkeley	0.49	0.36	0.41	0.53	0.30	0.39
sentimentr_sentiword	0.46	0.28	0.35	0.50	0.29	0.37
sentiment_stanford	0.55	0.37	0.45	0.45	0.28	0.35
sentiment_slangsd	0.57	0.35	0.44	0.41	0.26	0.32
sentiment_socal_google	0.58	0.23	0.33	0.54	0.21	0.30

Table 132 2016 Dover Micro/Macro/F-measure Precision and Recall (MR2)

Dictionary	Micro-Precisio	Micro-Recal	Micro-F-Sco	Macro-Precisio	Macro-Reca	Macro-F-Sco
sentiment_vadar	0.54	0.60	0.57	0.58	0.58	0.58
sentiment_jockers_rinker	0.45	0.61	0.52	0.52	0.63	0.57
sentimentr_jockers	0.46	0.60	0.52	0.53	0.61	0.57
syuzhet_jockers	0.45	0.60	0.51	0.52	0.62	0.56
syuzhet_afinn	0.54	0.61	0.58	0.54	0.56	0.55
sentimentr_huliu	0.59	0.58	0.58	0.57	0.51	0.54
syuzhet_bing	0.59	0.57	0.58	0.57	0.50	0.53
sentiment_sentistrength	0.54	0.62	0.58	0.50	0.55	0.52
sentiment_inquirer	0.61	0.55	0.58	0.56	0.47	0.51
sentiment_loughran_mcdonald	0.59	0.61	0.60	0.47	0.51	0.49
combined_dictionary	0.41	0.50	0.45	0.45	0.49	0.47
sentiment_senticnet	0.24	0.31	0.27	0.40	0.50	0.45
sentiment_nrc	0.49	0.40	0.44	0.50	0.39	0.44
syuzhet_nrc	0.50	0.41	0.45	0.51	0.38	0.44
sentiment_socal_google	0.57	0.42	0.48	0.51	0.34	0.41
sentimentr_sentiword	0.26	0.33	0.29	0.38	0.44	0.41
sentiment_stanford	0.42	0.53	0.47	0.34	0.42	0.38
sentiment_berkeley	0.25	0.25	0.25	0.40	0.35	0.37
sentiment_slangsd	0.38	0.53	0.44	0.30	0.42	0.35

Table 133 2016 MMM Micro/Macro/F-measure Precision and Recall (MR2)

Dictionary	Micro-Precisio	Micro-Recal	Micro-F-Sco	Macro-Precisio	Macro-Reca	Macro-F-Sco
sentiment_vadar	0.55	0.61	0.58	0.58	0.56	0.57
syuzhet_jockers	0.47	0.66	0.55	0.52	0.63	0.57
sentimentr_jockers	0.47	0.65	0.55	0.51	0.62	0.56
sentiment_jockers_rinker	0.46	0.66	0.54	0.50	0.63	0.56
syuzhet_afinn	0.56	0.64	0.60	0.54	0.55	0.55
sentimentr_huliu	0.61	0.60	0.61	0.57	0.51	0.53
syuzhet_bing	0.62	0.58	0.60	0.58	0.49	0.53
sentiment_inquirer	0.62	0.54	0.58	0.57	0.45	0.51
combined_dictionary	0.44	0.50	0.47	0.44	0.47	0.45
sentiment_loughran_mcdonald	0.59	0.59	0.59	0.44	0.46	0.45
sentiment_senticnet	0.26	0.32	0.29	0.40	0.47	0.44
sentimentr_sentiword	0.32	0.39	0.35	0.38	0.48	0.42
sentiment_nrc_cat	0.53	0.37	0.43	0.54	0.35	0.42
syuzhet_nrc	0.54	0.36	0.43	0.55	0.34	0.42
sentiment_berkeley	0.26	0.26	0.26	0.43	0.37	0.40
sentiment_socal_google	0.56	0.37	0.44	0.48	0.30	0.37
sentiment_slangsd	0.43	0.47	0.45	0.34	0.38	0.36
sentiment_stanford	0.45	0.50	0.47	0.34	0.37	0.36
sentiment_sentistrength	0.42	0.47	0.44	0.33	0.37	0.35

Table 134 2015 MMM Micro/Macro/F-measure Precision and Recall (MR2)

10.11.4.3 MR3 Results

Dictionaries	Micro-Precisi	Micro-Reca	Micro-F-Sco	Macro-Precisi	Macro-Recall	Macro-F-Sco
sentiment_jockers_rinker	0.19	0.53	0.28	0.26	0.55	0.35
sentimentr_jockers	0.20	0.54	0.29	0.27	0.55	0.36
sentimentr_huliu	0.40	0.72	0.51	0.29	0.47	0.36
sentimentr_sentiword	0.12	0.28	0.17	0.23	0.48	0.31
sentiment_nrc	0.40	0.50	0.44	0.36	0.40	0.38
sentiment_loughran_mcdonald	0.43	0.80	0.56	0.27	0.42	0.33
sentiment_senticnet	0.06	0.14	0.08	0.23	0.52	0.32
sentiment_inquirer	0.44	0.70	0.54	0.34	0.47	0.40
sentiment_slangsd	0.32	0.66	0.43	0.21	0.43	0.28
sentiment_socal_google	0.48	0.57	0.52	0.29	0.37	0.32
sentiment_vadar	0.29	0.62	0.40	0.32	0.50	0.39
sentiment_stanford	0.37	0.77	0.50	0.23	0.46	0.31
syuzhet_jockers	0.20	0.54	0.29	0.27	0.55	0.36
syuzhet_bing	0.42	0.72	0.53	0.32	0.47	0.38
syuzhet_afinn	0.32	0.69	0.44	0.29	0.52	0.37
syuzhet_nrc	0.41	0.51	0.45	0.35	0.40	0.37
sentiment_berkeley	0.07	0.12	0.09	0.25	0.44	0.32
sentiment_sentistrength	0.34	0.70	0.46	0.21	0.46	0.29
combined_dictionary	0.25	0.53	0.34	0.23	0.46	0.31

Table 135 2015 MMM Micro/Macro/F-measure Precision and Recall (MR3)

Dictionaries	Micro-Precisio	Micro-Recal	Micro-F-Sco	Macro-Precisio	Macro-Recall	Macro-F-Sco
sentiment_jockers_rinker	0.21	0.48	0.29	0.41	0.58	0.48
sentimentr_jockers	0.22	0.49	0.30	0.41	0.60	0.49
sentimentr_huliu	0.44	0.72	0.55	0.48	0.56	0.52
sentimentr_sentiword	0.10	0.21	0.14	0.27	0.50	0.35
sentiment_nrc	0.37	0.50	0.43	0.46	0.49	0.48
sentiment_loughran_mcdonald	0.50	0.85	0.63	0.27	0.42	0.33
sentiment_senticnet	0.04	0.10	0.06	0.33	0.55	0.41
sentiment_inquirer	0.49	0.73	0.58	0.41	0.50	0.45
sentiment_slangsd	0.32	0.77	0.45	0.18	0.46	0.26
sentiment_socal_google	0.50	0.60	0.55	0.37	0.43	0.40
sentiment_vadar	0.31	0.59	0.41	0.45	0.62	0.52
sentiment_stanford	0.38	0.80	0.51	0.26	0.49	0.34
syuzhet_jockers	0.21	0.49	0.30	0.38	0.59	0.47
syuzhet_bing	0.45	0.72	0.56	0.40	0.57	0.47
syuzhet_afinn	0.35	0.66	0.45	0.39	0.57	0.46
syuzhet_nrc	0.38	0.52	0.44	0.44	0.49	0.46
sentiment_berkeley	0.11	0.17	0.13	0.28	0.54	0.37
sentiment_sentistrength	0.42	0.80	0.55	0.42	0.64	0.51
combined_dictionary	0.26	0.54	0.35	0.35	0.57	0.43

Table 136 2016 MMM Micro/Macro/F-measure Precision and Recall (MR3)

Dictionaries	Micro-Precisio	Micro-Recal	Micro-F-Sco	Macro-Precisio	Macro-Recall	Macro-F-Sco
sentiment_jockers_rinker	0.11	0.36	0.17	0.31	0.56	0.40
sentimentr_jockers	0.12	0.37	0.18	0.33	0.57	0.42
sentimentr_huliu	0.24	0.57	0.34	0.34	0.54	0.42
sentimentr_sentiword	0.10	0.20	0.13	0.27	0.50	0.35
sentiment_nrc	0.31	0.58	0.40	0.28	0.48	0.35
sentiment_loughran_mcdonald	0.44	0.78	0.56	0.28	0.40	0.33
sentiment_senticnet	0.05	0.11	0.07	0.35	0.48	0.40
sentiment_inquirer	0.33	0.57	0.42	0.32	0.45	0.38
sentiment_slangsd	0.33	0.68	0.44	0.19	0.47	0.27
sentiment_socal_google	0.38	0.52	0.44	0.42	0.42	0.42
sentiment_vadar	0.15	0.45	0.23	0.30	0.57	0.39
sentiment_stanford	0.31	0.73	0.44	0.29	0.52	0.37
syuzhet_jockers	0.11	0.36	0.17	0.31	0.57	0.40
syuzhet_bing	0.28	0.61	0.38	0.38	0.54	0.45
syuzhet_afinn	0.19	0.57	0.29	0.27	0.57	0.37
syuzhet_nrc	0.32	0.58	0.41	0.29	0.48	0.36
sentiment_berkeley	0.07	0.18	0.10	0.26	0.49	0.34
sentiment_sentistrength	0.22	0.67	0.33	0.27	0.59	0.37
combined_dictionary	0.20	0.53	0.29	0.32	0.52	0.40

Table 137 2016 Dover Micro/Macro/F-measure Precision and Recall (MR3)

Dictionaries	Micro-Precision	Micro-Reca	Micro-F-Sco	Macro-Precis	Macro-Reca	Macro-F-Sco
sentiment_jockers_rinker	0.23	0.44	0.30	0.40	0.55	0.46
sentimentr_jockers	0.24	0.45	0.31	0.41	0.55	0.47
sentimentr_huliu	0.42	0.66	0.52	0.46	0.54	0.49
sentimentr_sentiword	0.11	0.21	0.14	0.30	0.50	0.38
sentiment_nrc	0.36	0.55	0.44	0.37	0.44	0.40
sentiment_loughran_mcdonald	0.57	0.82	0.67	0.35	0.41	0.38
sentiment_senticnet	0.06	0.11	0.08	0.31	0.53	0.39
sentiment_inquirer	0.45	0.63	0.52	0.42	0.40	0.41
sentiment_slangsd	0.31	0.72	0.43	0.18	0.44	0.25
sentiment_socal_google	0.47	0.55	0.51	0.35	0.36	0.35
sentiment_vadar	0.31	0.51	0.38	0.43	0.52	0.47
sentiment_stanford	0.35	0.77	0.48	0.34	0.61	0.44
syuzhet_jockers	0.24	0.45	0.31	0.41	0.56	0.47
syuzhet_bing	0.44	0.68	0.54	0.44	0.54	0.49
syuzhet_afinn	0.35	0.58	0.44	0.43	0.57	0.49
syuzhet_nrc	0.37	0.55	0.44	0.38	0.43	0.40
sentiment_berkeley	0.07	0.12	0.09	0.31	0.43	0.36
sentiment_sentistrength	0.39	0.74	0.51	0.40	0.55	0.46
combined_dictionary	0.27	0.46	0.34	0.38	0.52	0.44

Table 138 2016 AA Micro/Macro/F-measure Precision and Recall (MR3)

10.12 Dictionary Approach Results

10.12.1 MR1 Results

10.12.1.1 MR1 Model Results

2015 MMM Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**Naïve Bayes (CARET)**
Negative	184	195	194	190	207	224
Neutral	108	105	101	102	93	76
Positive	8	NA	5	8	NA	0

Table 139 Dictionary Approach - MR1 2015 MMM Model Results

2016 MMM Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**Naïve Bayes (CARET)**
Negative	128	155	157	151	145	165
Neutral	148	133	125	137	134	106
Positive	24	12	18	12	21	29

Table 140 Dictionary Approach - MR1 2016 MMM Model Results

2016 Dover Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**Naïve Bayes (CARET)**
Negative	281	281	276	268	265	300
Neutral	19	19	24	32	35	NA
Positive	NA	NA	NA	NA	NA	NA

Table 141 Dictionary Approach - MR1 2016 Dover Model Results

2016 Anti-Austerity Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**Naïve Bayes (CARET)**
Negative	158	169	165	170	157	154
Neutral	124	116	120	111	125	107
Positive	18	15	15	19	18	39

Table 142 Dictionary Approach - MR1 2016 Anti-Austerity Model Results

2015 MMM Machine Learning Results Version 1 Cut Off 0.9						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**Naïve Bayes (CARET)**
Negative	6766	7525	7327	7383	8198	9784
Neutral	4857	4731	4499	4551	4058	2305
Positive	633	NA	430	322	NA	167

Table 143 Dictionary Approach - MR1 Grouped Model Results

MR1 Grouped Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**Naïve Bayes (CARET)**
Negative	727	765	724	743	662	895
Neutral	404	384	407	402	460	245
Positive	69	51	69	55	78	60

Table 144 Dictionary Approach - MR1 Grouped Model Results

10.12.2 MR2 Results

10.12.2.1 MR2 Model Results

2015 MMM Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**Naïve Bayes (CARET)**
Negative	110	123	136	128	173	146
Neutral	188	177	162	168	127	154
Positive	2	NA	2	4	NA	NA

Table 145 MR2 2015 MMM Model Results

2016 MMM Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**Naïve Bayes (CARET)**
Negative	49	52	68	60	276	55
Neutral	229	241	214	223	24	216
Positive	22	7	18	17	NA	32

Table 146 MR2 2016 MMM Model Results

2016 Dover Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**Naïve Bayes (CARET)**
Negative	237	290	251	237	261	300
Neutral	63	10	59	60	39	NA
Positive	NA	NA	NA	3	NA	NA

Table 147 MR2 2016 Dover Model Results

2016 Anti-Austerity Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**Naïve Bayes (CARET)**
Negative	NA	NA	13	18	NA	NA
Neutral	299	300	284	270	300	284
Positive	1	NA	3	12	NA	16

Table 148 MR2 2016 Anti-Austerity Model Results

2016 MR2 Grouped Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**Naïve Bayes (CARET)**
Negative	341	428	450	451	601	567
Neutral	843	745	723	722	599	584
Positive	16	27	27	27	NA	49

Table 149 MR2 Grouped Model Results

10.12.3 MR1 and MR2 Results

10.12.3.1 MR1 and MR2 Model Results

2015 MMM Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**Naïve Bayes (CARET)**
Negative	87	94	87	91	107	104
Neutral	77	74	78	74	61	64
Positive	4	NA	3	3	NA	NA

Table 150 Agreed 2015 MMM Model Results

2016 MMM Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**Naïve Bayes (CARET)**
Negative	62	58	68	64	94	50
Neutral	92	95	79	85	60	96
Positive	14	15	21	19	14	22

Table 151 Agreed 2016 MMM Model Results

2016 Dover Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**Naïve Bayes (CARET)**
Negative	156	152	149	148	151	168
Neutral	11	16	18	20	17	NA
Positive	1	NA	1	NA	NA	NA

Table 152 Agreed 2016 Dover Model Results

2016 Anti-Austerity Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**Naïve Bayes (CARET)**
Negative	34	41	44	44	43	22
Neutral	122	121	116	116	104	119
Positive	12	6	8	8	21	27

Table 153 Agreed 2016 Anti-Austerity Model Results

2016 MR1 and MR2: Grouped Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**Naïve Bayes (CARET)**
Negative	292	327	329	315	326	355
Neutral	358	323	316	331	315	270
Positive	22	22	27	26	31	47

Table 154 Agreed MR1 & MR2 Grouped Model Results

10.12.4 MR1 SOMEWHATS Results

	MMM 2015				MMM 2016				DOVER 2016				Anti-Austerity 2016			
Sentiment Category	**MR1& MR2**	**MR1& MR3**	**MR2& MR3**	**All**	**MR1& MR2**	**MR1& MR3**	**MR2& MR3**	**All**	**MR1& MR2**	**MR1& MR3**	**MR2& MR3**	**All**	**MR1& MR2**	**MR1& MR3**	**MR2& MR3**	**All**
Strongly Negative	494	75	56	496	369	29	25	369	936	83	77	937	185	40	29	198
Somewhat Negative	354	750	487	21	315	648	383	1	294	1132	888	5	514	674	183	0
Total Negative	**848**	**825**	**543**	**517**	**684**	**677**	**408**	**370**	**1230**	**1215**	**965**	**942**	**699**	**714**	**212**	**198**
Strongly Positive	54	13	16	58	88	10	11	88	15	5	5	19	74	39	38	98
Somewhat Positive	49	78	84	7	96	116	166	3	36	30	56	2	115	119	141	1
Total Positive	**103**	**91**	**100**	**65**	**184**	**126**	**177**	**91**	**51**	**35**	**61**	**21**	**189**	**158**	**179**	**99**
Strongly Neutral	NA	NA	NA	508	NA	NA	NA	600	NA	NA	NA	167	NA	NA	NA	572
Somewhat Neutral	NA	NA	NA	410	NA	NA	NA	439	NA	NA	NA	370	NA	NA	NA	631
Neutral	549	584	857	NA	632	697	915	NA	219	250	474	NA	612	628	1109	NA
Total Neutral	**549**	**584**	**857**	**918**	**632**	**697**	**915**	**1039**	**219**	**250**	**474**	**537**	**612**	**628**	**1109**	**1203**
Total	**1500**	**1500**	**1500**	**1500**	**1500**	**1500**	**1500**	**1500**	**1500**	**1500**	**1500**	**1500**	**1500**	**1500**	**1500**	**1500**
Proportion Negative	56.53	55.00	36.20	34.47	45.60	45.13	27.20	24.67	82.00	81.00	64.33	62.80	46.60	47.60	14.13	13.20
Proportion Positive	6.87	6.07	6.67	4.33	12.27	8.40	11.80	6.07	3.40	2.33	4.07	1.40	12.60	10.53	11.93	6.60
Proportion Neutral	36.60	38.93	57.13	61.20	42.13	46.47	61.00	69.27	14.60	16.67	31.60	35.80	40.80	41.87	73.93	80.20
Total	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

Table 155 MR1 SomeWhats Proportional Results

Note: If a human annotator agrees then it is strongly positive/neutral/negative and if annotator disagrees then declared a somewhat positive/neutral/negative. The f-measure results for MR1 and MR2 SOMEWHATS produced very poor results. The increase from 3 to 5 categories seems to have led to the poor results.

10.13 Machine Learning Approach

10.13.1 MR1 Results

10.13.1.1 MR1 Model Results

2015 MMM Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	166	166	144	130	241	145
Neutral	119	124	155	168	57	137
Positive	15	10	1	2	2	18

Table 156 MR1 2015 MMM Model Train Results

2016 MMM Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	129	152	96	91	61	66
Neutral	147	133	193	194	227	187
Positive	24	15	11	15	12	47

Table 157 MR1 2016 MMM Model Train Results

2016 Dover Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	282	286	292	289	293	278
Neutral	15	12	7	7	6	17
Positive	3	2	1	4	1	5

Table 158 MR1 2016 Dover Model Train Results

2016 Anti-Austerity Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	166	185	218	237	260	160
Neutral	109	100	76	60	40	111
Positive	25	15	6	3	NA	29

Table 159 MR1 2016 Anti-Austerity Model Train Results

10.13.1.2 MR1 Test Results

2015 MMM Test Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	14526	15030	8399	5236	27485	13916
Neutral	11340	11841	20599	23640	1394	11078
Positive	3554	2549	422	544	541	4426

Table 160 MR1 2015 MMM Test Results

2016 MMM Test Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	6607	6830	3467	3119	1276	3162
Neutral	6998	7661	11500	11787	13732	8761
Positive	1886	1000	524	585	483	3568

Table 161 MR1 2016 MMM Test Results

2016 Dover Test Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	2709	2784	2985	3018	3028	2730
Neutral	392	337	147	114	104	294
Positive	73	53	42	42	42	150

Table 162 MR1 2016 Dover Test Results

2016 Anti-Austerity Test Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	13770	14716	19805	24045	25743	14931
Neutral	11752	12086	8511	4919	3529	9731
Positive	4441	3161	1647	999	691	5301

Table 163 MR1 2016 Anti-Austerity Test Results

10.13.2 MR2 Results

	MMM 2015					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	42	50	0	52	38	2
Neutral	14	174	2	44	130	16
Positive	2	16	0	7	8	3
Precision	0.48			0.46		
Recall	0.46			0.4733333		
F1 Score	0.4566667			0.4633333		
Accuracy	0.72			0.6166667		
	Support Vector Machine			**Bagging**		
Negative	47	45	0	42	48	2
Neutral	26	159	5	14	175	1
Positive	1	16	1	1	17	0
Precision	0.51			0.49		
Recall	0.47			0.46		
F1 Score	0.48			0.46		
Accuracy	0.69			0.7233333		
	Tree			**NNETWORK**		
Negative	42	49	1	26	59	7
Neutral	33	156	1	37	145	8
Positive	3	15	0	6	9	3
Precision	0.4166667			0.41		
Recall	0.4266667			0.4033333		
F1 Score	0.42			0.4033333		
Accuracy	0.66			0.58		

Table 164 2015 MMM results for machine learning algorithms

	MMM 2016					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	16	67	0	31	47	5
Neutral	31	142	0	61	103	10
Positive	11	31	1	11	26	6
Precision	0.4566667			0.3933333		
Recall	0.3433333			0.3666667		
F1 Score	0.32			0.37		
Accuracy	0.53			0.4666667		
	Support Vector Machine			**Bagging**		
Negative	18	65	0	18	65	0
Neutral	45	127	2	28	144	2
Positive	11	28	4	11	31	1
Precision	0.4966667			0.4166667		
Recall	0.3466667			0.3566667		
F1 Score	0.3466667			0.3333333		
Accuracy	0.4966667			0.5433333		
	Tree			**NNETWORK**		
Negative	27	56	0	15	64	4
Neutral	42	131	1	46	120	8
Positive	9	33	1	8	29	6
Precision	0.4833333			0.37		
Recall	0.3666667			0.3366667		
F1 Score	0.35			0.34		
Accuracy	0.53			0.47		

Table 165 2016 MMM results for machine learning algorithms

	Dover 2016					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	203	4	0	167	36	4
Neutral	75	13	1	50	34	5
Positive	4	0	0	3	0	1
Precision	0.4933333			0.45		
Recall	0.3766667			0.48		
F1 Score	0.36			0.45		
Accuracy	0.72			0.6733333		
	Support Vector Machine			**Bagging**		
Negative	193	14	0	203	3	1
Neutral	61	27	1	81	7	1
Positive	2	1	1	4	0	0
Precision	0.63			0.4666667		
Recall	0.4933333			0.3533333		
F1 Score	0.5233333			0.32		
Accuracy	0.7366667			0.70		
	Tree			**NNETWORK**		
Negative	207	0	0	170	33	4
Neutral	85	3	1	57	25	7
Positive	4	0	0	0	3	1
Precision	0.5666667			0.4133333		
Recall	0.3433333			0.45		
F1 Score	0.2933333			0.41		
Accuracy	0.70			0.6533333		

Table 166 2016 Dover results for machine learning algorithms

	Anti-Austerity 2016					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	5	35	0	15	23	2
Neutral	2	236	1	23	199	17
Positive	0	17	4	0	12	9
Precision	0.7766667			0.52		
Recall	0.4333333			0.5466667		
F1 Score	0.4733333			0.53		
Accuracy	0.8166667			0.7433333		
	Support Vector Machine			**Bagging**		
Negative	5	35	0	5	35	0
Neutral	5	232	2	2	234	3
Positive	1	17	3	1	16	4
Precision	0.6233333			0.67		
Recall	0.41			0.43		
F1 Score	0.4366667			0.4566667		
Accuracy	0.80			0.81		
	Tree			**NNETWORK**		
Negative	5	35	0	5	34	1
Neutral	7	228	4	19	208	12
Positive	1	17	3	3	12	6
Precision	0.54			0.4433333		
Recall	0.4033333			0.4266667		
F1 Score	0.42			0.43		
Accuracy	0.7866667			0.73		

Table 167 2016 Anti-Austerity results for machine learning algorithms

	MR2 Grouped					
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
	Forest			**Max Entropy**		
Negative	259	151	1	250	139	22
Neutral	97	577	13	155	473	59
Positive	7	71	24	18	53	31
Precision	0.6866667			0.5266667		
Recall	0.57			0.5333333		
F1 Score	0.60			0.53		
Accuracy	0.7166667			0.6283333		
	Support Vector Machine			**Bagging**		
Negative	233	177	1	191	217	3
Neutral	90	581	16	92	588	7
Positive	2	75	25	6	79	17
Precision	0.6733333			0.6533333		
Recall	0.5566667			0.4966667		
F1 Score	0.5866667			0.52		
Accuracy	0.6991667			0.6633333		
	Tree			**NNETWORK**		
Negative	123	288	0	99	216	96
Neutral	85	602	0	73	513	101
Positive	6	96	0	16	46	40
Precision	0.3933333			0.4533333		
Recall	0.3933333			0.46		
F1 Score	0.37			0.4233333		
Accuracy	0.6041667			0.5433333		

Table 168 MR2 Grouped results for machine learning algorithms

10.13.2.1 MR2 Model Results

2015 MMM Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	103	74	58	57	78	69
Neutral	176	220	240	240	220	213
Positive	21	6	2	3	2	18

Table 169 MR2 2015 MMM Model Train Results

2016 MMM Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	103	74	58	57	78	69
Neutral	176	220	240	240	220	213
Positive	21	6	2	3	2	18

Table 170 MR2 2016 MMM Model Train Results

2016 Dover Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	220	256	282	288	296	227
Neutral	70	42	17	10	3	61
Positive	10	2	1	2	1	12

Table 171 MR2 2016 Dover Model Train Results

2016 Anti-Austerity Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	38	11	7	8	13	27
Neutral	234	284	288	285	280	254
Positive	28	5	5	7	7	19

Table 172 MR2 2016 Anti-Austerity Model Train Results

MR2 Grouped Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	423	325	363	289	214	188
Neutral	665	833	799	884	986	775
Positive	112	42	38	27	NA	237

Table 173 MR2 Grouped Model Train Results

10.13.2.2 MR2 Test Results

2015 MMM Test Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	9852	6754	3649	2316	2722	6999
Neutral	14323	20433	25441	26641	26157	17634
Positive	5245	2233	330	463	541	4787

Table 174 MR2 2015 MMM Test Results

2016 MMM Test Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	3912	2373	1285	1186	1244	3645
Neutral	9408	11955	13619	13843	13890	9033
Positive	2171	1163	587	462	357	2813

Table 175 MR2 2016 MMM Test Results

2016 Dover Test Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	2096	2188	2697	3030	3098	2239
Neutral	938	958	443	116	41	688
Positive	140	28	34	28	35	247

Table 176 MR2 2016 Dover Test Results

2016 Anti-Austerity Test Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	3713	1745	332	325	1428	4183
Neutral	21220	26751	28911	28777	27720	21162
Positive	5030	1467	720	861	815	4618

Table 177 MR2 2016 Anti-Austerity Test Results

MR2 Grouped Test Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	24158	14650	12387	7799	4861	9500
Neutral	44540	60343	62956	68467	73187	50783
Positive	9350	3055	2705	1782	NA	17765

Table 178 MR2 Grouped Test Results

10.13.3 MR1 & MR2 Results

10.13.3.1 MR1 & MR2 Model Results

2015 MMM Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	96	97	74	68	63	64
Neutral	63	66	91	96	103	94
Positive	9	5	3	4	2	10

Table 179 MR1 & MR2 2015 MMM Model Train Results

2016 MMM Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	66	61	48	56	56	50
Neutral	82	89	110	103	103	78
Positive	20	18	10	9	9	40

Table 180 MR1 & MR2 2016 MMM Model Train Results

2016 Dover Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	148	151	160	157	154	144
Neutral	18	17	8	9	14	22
Positive	2	NA	NA	2	NA	2

Table 181 MR1 & MR2 2016 Dover Model Train Results

2016 Anti-Austerity Model Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	50	38	18	16	15	21
Neutral	111	123	145	146	148	122
Positive	7	7	5	6	5	25

Table 182 MR1 & MR2 2016 Anti-Austerity Model Train Results

10.13.3.2 MR1 & MR2 Test Results

2015 MMM Test Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	14900	12311	4337	3797	3339	9706
Neutral	10585	14476	24514	24853	25538	14400
Positive	3935	2633	569	770	543	5314

Table 183 MR1 & MR2 2015 MMM Test Results

2016 MMM Test Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	6034	4460	1483	1633	1402	4474
Neutral	8176	9870	13646	13417	13728	8991
Positive	1281	1161	362	441	361	2026

Table 184 MR1 & MR2 2016 MMM Test Results

2016 Dover Test Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	2698	2708	3034	3008	2978	2703
Neutral	396	417	123	132	196	374
Positive	80	49	17	34	NA	97

Table 185 MR1 & MR2 2014 Dover Test Results

2016 Anti-Austerity Test Results						
SENTIMENT LABEL	**MAXENTROPY LABEL**	**SVM LABEL**	**FORESTS LABEL**	**BAGGING LABEL**	**TREE LABEL**	**NNETWORK LABEL**
Negative	8598	6668	1855	1935	2405	6822
Neutral	17636	19705	26937	26233	25148	17761
Positive	3729	3590	1171	1795	2410	5380

Table 186 MR1 & MR2 2016 Anti-Austerity Test Results

10.14 Gold Standard Human Annotation Agreement Results

MMM 2015				
Category	**1st rat**	**2nd rat**	**Percentage agreement (Tolerance=0)**	**Cohen's Kappa for 2 Raters (Weights: unweighted)**
Counted Positive	42.00	158.00	Subjects = 840	Subjects = 840
Counted Negative	385.00	340.00	Raters = 2	Raters = 2
Counted Neutral	413.00	342.00	%-agree = 71.7	Kappa = 0.532
Total	840.00	840.00		z = 20.4
				p-value = 0
Positive Proportion	5.00	18.81	**Cohen's Kappa for 2 Raters (Weights: equal)**	**Cohen's Kappa for 2 Raters (Weights: squared)**
Negative Proportion	45.83	40.48	Subjects = 840	Subjects = 840
Neutral Proportion	49.17	40.71	Raters = 2	Raters = 2
Total	100.00	100.00	Kappa = 0.496	Kappa = 0.45
			z = 19.3	z = 14
Matched	602.00		p-value = 0	p-value = 0
Unmatched	238.00		**Krippendorff's alpha**	
Total	840.00		Subjects = 840	
			Raters = 2	
Proportion Matched	71.67		alpha = 0.495	
Proportion Unmatched	28.33			

Table 187 2015 MMM Gold Standard and Majority Voting Agreement Level

MMM 2016				
Category	**1st rat**	**2nd rat**	**Percentage agreement (Tolerance=0)**	**Cohen's Kappa for 2 Raters (Weights: unweighted)**
Counted Positive	68.00	202.00	Subjects = 840	Subjects = 840
Counted Negative	287.00	255.00	Raters = 2	Raters = 2
Counted Neutral	485.00	383.00	%-agree = 66.8	Kappa = 0.459
Total	840.00	840.00		z = 18.7
				p-value = 0
Positive Proportion	8.10	24.05	**Cohen's Kappa for 2 Raters (Weights: equal)**	**Cohen's Kappa for 2 Raters (Weights: squared)**
Negative Proportion	34.17	30.36	Subjects = 840	Subjects = 840
Neutral Proportion	57.74	45.60	Raters = 2	Raters = 2
Total	100.00	100.00	Kappa = 0.448	Kappa = 0.434
			z = 18.3	z = 13.4
Matched	561.00		p-value = 0	p-value = 0
Unmatched	279.00		**Krippendorff's alpha**	
Total	840.00		Subjects = 840	
			Raters = 2	
Proportion Matched	66.79		alpha = 0.435	
Proportion Unmatched	33.21			

Table 188 2016 MMM Gold Standard and Majority Voting Agreement Level

Dover 2016				
Category	**1st rat**	**2nd rat**	**Percentage agreement (Tolerance=0)**	**Cohen's Kappa for 2 Raters (Weights: unweighted)**
Counted Positive	12.00	186.00	Subjects = 840	Subjects = 840
Counted Negative	693.00	492.00	Raters = 2	Raters = 2
Counted Neutral	135.00	162.00	%-agree = 53.9	Kappa = 0.0454
Total	840.00	840.00		z = 2.1
				p-value = 0.0359
Positive Proportion	1.43	22.14	**Cohen's Kappa for 2 Raters (Weights: equal)**	**Cohen's Kappa for 2 Raters (Weights: squared)**
Negative Proportion	82.50	58.57	Subjects = 840	Subjects = 840
Neutral Proportion	16.07	19.29	Raters = 2	Raters = 2
Total	100.00	100.00	Kappa = 0.0369	Kappa = 0.0291
			z = 1.79	z = 1.27
Matched	453.00		p-value = 0.0735	p-value = 0.203
Unmatched	387.00		**Krippendorff's alpha**	
Total	840.00		Subjects = 840	
			Raters = 2	
Proportion Matched	53.93		alpha = -0.0384	
Proportion Unmatched	46.07			

Table 189 2016 Dover Gold Standard and Majority Voting Agreement Level

Aniti-Austerity 2016				
Category	**1st rat**	**2nd rat**	**Percentage agreement (Tolerance=0)**	**Cohen's Kappa for 2 Raters (Weights: unweighted)**
Counted Positive	74.00	287.00	Subjects = 840	Subjects = 840
Counted Negative	185.00	184.00	Raters = 2	Raters = 2
Counted Neutral	581.00	369.00	%-agree = 61.1	Kappa = 0.37
Total	840.00	840.00		z = 17.1
				p-value = 0
Positive Proportion	8.81	34.17	**Cohen's Kappa for 2 Raters (Weights: equal)**	**Cohen's Kappa for 2 Raters (Weights: squared)**
Negative Proportion	22.02	21.90	Subjects = 840	Subjects = 840
Neutral Proportion	69.17	43.93	Raters = 2	Raters = 2
Total	100.00	100.00	Kappa = 0.372	Kappa = 0.374
			z = 17.1	z = 12.3
Matched	513.00		p-value = 0	p-value = 0
Unmatched	327.00		**Krippendorff's alpha**	
Total	840.00		Subjects = 840	
			Raters = 2	
Proportion Matched	61.07		alpha = 0.352	
Proportion Unmatched	38.93			

Table 190 2016 Anti-Austerity Gold Standard and Majority Voting Agreement Level