RESEARCH

Comparative Analysis of Lexicon-Based Sentiment Analysis Methods

Dr James Baldwin[1*], Dr Teresa Brunsdon[2], Dr Jotham Gaudoin[3] and Dr Laurence Hirsch[4]

*Correspondence:
j.baldwin@shu.ac.uk
teresa.brunsdon@warwick.ac.uk
jotham.gaudoin@open.ac.uk
l.hirsch@shu.ac.uk

Full list of author information is available at the end of the article

Abstract

Sentiment Analysis studies the opinions, sentiments and emotions expressed at sentence or document level. Machine learning and lexicon-based approaches have been successfully used to achieve this. This paper will focus on the lexicon-based approach. In contrast to most existing research, we compare the effectiveness of multiple dictionaries across a series of datasets related to public order events. The comparison will look to understand the possible benefits and limitations of sentiment analysis methods, which will bench marked against each other in their evaluation of results. The evaluation will be based four labelled datasets, covering messages on posts related to public order events. The results will highlight the extent of how well each of these methods perform across the datasets comparing sentence level analysis with range of sentiment analysis techniques.

Keywords: sentiment analysis; social media; lexicon approach

1 Introduction

There is a wealth of research conducted in sentiment analysis in the last several years, and continues to develop in many topical domains on the web and social media, where thoughts, opinions and/ or attitudes on data that has been used a range of areas, such as innovation of products. Sentiment Analysis identifies and measures whether the text being analysed is positive, negative or neutral as an entity, such as people, organisation, event, location, or a topic. 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 [1].

Organisations now have social media teams to monitor events and actively release information, quickly reacting to situations of widespread interest [1]. 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, monitoring brand reputation, reviews of products/ services, and marketing. There is wide interest on how to apply sentiment analysis, which has led to continued development of lexicons, machine learning process and addressing other factors, such as negation to enhance the accuracy of the outcome. The lexical methods vary based on the topic domain it is applied to and can be impacted in time where use of language and/ or meaning of words change in time. For instance, AFINN was developed work on financial reports, VADAR for social network data, SentiStrength based on short informal messages related to social network sites, blogs and discussion forums, and Syuzhet is designed for literature in humanities. These techniques are widely used in various ways within the research community, and have been applied in a range of publications, where applied in original form and/ or been adapted to suit the purpose a research project. For example, SentiStrength was used to measure positivity and negativity of online news, and VADER to study patterns of smoking and drinking abstinence in social media [8]. As SentiBench [8] suggested there is no state of the art established as "researchers tend to accept any popular method as a valid methodology" to analyse sentiment which seemingly indicates less is known about their performance of the lexicon(s) on a wide range of datasets. This suggests there is a need for further investigation of a thorough comparison of sentiment analysis methods on multiple datasets to further understand their suitability for their application. We will provide a thorough benchmark of comparison of 19 lexicon dictionaries based on 4 different Twitter datasets where the focus will be on analysing the sentiment at sentence level for the tweets. We conducted a wider study that was based on public order context, which is why the nature of the data we focus on in this publication is on public order events.

The experimental results provided a series of important findings, for instance, we show that there are specific dictionaries e.g., Jockers series (refer to section 4) and VADAR that achieve best prediction on the Twitter data events, and also find commonality with other publications (refer to section 4.1) which indicate a series of common dictionaries that have performed at a top level consistently across some different datasets. However, there are some dictionaries in the context of the public order events based on Twitter data that have performed less well e.g., Slangsd and Socal Google which are geared more towards specific contexts based on how they were designed in the first place (refer to section 2.3 Table 1). This demonstrates that existing lexicons vary regarding their agreement as the same content could be interpreted quite differently depending on the choice of a sentiment method. We noted that most methods are more accurate in correctly classifying negative than positive text, suggesting that current approaches tend to be biased in their analysis towards negativity. However, the balance of tweets some of the public order events were more negative in discussion except for Anti-Austerity which tended to show more positivity which why there is perhaps more bias to negativity. In this study, we quantify the prediction performance based on the 19 dictionaries we applied and compare this with existing efforts in the field across different types of datasets to identify commonality and differences between the lexicon's performance.

Based on these observations, our final contribution consists of the release of our combined dictionary, benefits and limitations of new combined dictionary and the comparison of sentiment analysis methods. This may help researchers and practitioners towards the further development of this research field. In the following sections we focus on existing position of sentiment analysis, comparing sentiment analysis methods, adaptation of method, and then move onto what framework that was applied in the project. We then discuss the findings and results of the lexicon approach, and compare this with other relevant research in this area to identify if there are commonality and/ or differences in their outcomes. Finally, conclude the paper linked with discussion on future work.

2 Background and Related Work

In the following section, we will discuss a series of important definitions and justify the benchmark comparison, and include literature survey that emphasise on different use cases where sentiment analysis has been applied and compare them.

2.1 Sentiment Analysis Approach

Sentiment analysis can be applied to various tasks, but we focus on comparing polarity of short text on a sentence-level [2]. The detection of polarity is common across sentiment methods that provides important insights to a series of different applications, social media is one that can be commonly sourced. Sentence-level sentiment analysis can be performed with supervision or not. A supervised learning approach has a benefit of being able to adapt and create trained models for different context for a purpose. A limitation of supervised method is the need of labelled data, which can be resource intensive. A lexical based approach is where has a pre-defined list of words, where each word is assigned a polarity score, but the lexical method may vary on their output dependent on the context on how they were created [2]. For instance, VADAR lexicon was to discover patterns of smoking and drinking abstinence within social media data [8]. Thus, different dictionaries have been created for a range of purposes and contexts, but it is challenging to create a unique lexical dictionary for different contexts and also requires linguist specialists' expertise. There are many dictionaries (e.g., SentiWordNet, SenticNet, Stanford and SentiStrength) that are based on the English language, but most are American English rather than UK English [1]. Furthermore, other dictionaries with different languages are sparse in comparison with English based dictionaries [2]. 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 [2].

There are series of sentiment analysis approaches which includes the machine learning approach, lexicon-based approach and hybrid approach. Our aim is to classify tweets as positive, negative and neutral, but to do this in a highly automated way to adapt to the high volume of social media data. We have chosen to adopt the hybrid approach, combining both lexicon and machine learning approaches to apply sentiment analysis to four Twitter datasets [2]. The lexicon-based approach will perform at sentence level to determine the polarity from the predefined dictionary while the machine learning algorithms including Support Vector Machine (SVM), Na¨ıve Bayes and Maximum Entropy will train a classifier by using the polarity for each sentence as determined by the lexicon [2]. 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 [2].

We adopt the hybrid approach where both a dictionary and machine learning approaches are applied and compare their performance, but for the focus on this paper we will focus on the lexicon approach.

2.2 Existing Sentiment Analysis Methods Research

There a large number of existing sentiment analysis approaches used in a variety of research projects, but as we observed in align with [8] there is "limited number of them have performed comparison among sentiment analysis methods" with highly specific datasets. The machine learning and lexicon-based approach has been developing in parallel, and there are some studies [7, 8, 3] that draw comparison on each approach. However, when reviewing a series of papers [9, 5, 10], it can be difficult to compare, as papers refer to accuracy or correlation of sentiment scores, but F1 measure is known as more of a reliable indicator to compare outcomes. Therefore, it can make it difficult to compare the results if similar indicators are not applied to determine strength of results.

In the existing research, there are introduction of new dictionaries [4], the comparison of dictionaries seem to be one or few lexicons, use of different datasets for evaluation. This can be challenging to draw comparisons of dictionaries to identify best method of approach in a universal way and for specific scenarios. Seemingly there is one paper that has carried out a thorough benchmark [8] that have compared 24 dictionaries of free/ paid against a series of datasets. This paper used 24 dictionaries, and used gold standard datasets in comments on BBC, Digg, NYT, TED and YouTube, movie reviews, Amazon products reviews and social network data based on Twitter topics. We will benchmark against the Twitter (social network data), as our datasets is based on four demonstration events on Twitter [8]. Also, this paper has focused on the use off-the-shelf tools that have been "extensively and recently used" which includes multiple commercial options with some free options. The focus of our paper is only on freely available lexicons. To best of our knowledge this paper has provided the only extensive bench-marking for the lexicon-based approach, therefore, this demonstrates a greater need for further work in the comparison of dictionaries with different datasets to understand which are best suited for specific or more generalised topic areas.

2.3 Applied Sentiment Analysis Methods

In this section, there is a description of the 18 lexicons with addition of 1 more where we created a combined dictionary of 11 lexicons (refer to section 2.5). The 18 lexicons that were identified with numerous research papers based on sentiment analysis and also within the programming language libraries that included lexicons within the package as well. A couple of the methods were made available to download on the Web or kindly shared by the authors, for example, SentiStrength was available to download.

Table 1 presents a series of methods with a description of each one, and the lexicon employed, number of terms used, their outputs (e.g., -1,0,1, meaning negative, neutral, and positive, respectively) and notes to provide some context for each lexicon package. 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 [2]. 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) [2]. The 18 dictionaries chosen are freely available, but SentiStrength is paid, however, is free of charge academic license. Other lexicons are available, such as LIWC, but this can only be applied as free trial and after that need to pay a fee, but the focus of the project targeted freely available option rather than paid options. The authors cited in Table 1 that created the dictionaries are highly respected in the field and are widely used in research where examples of some of these dictionaries are cited after Table 1.

Table 1 Range of Sentiment Analysis Lexicons

Lexicon name	Number of terms	Sentiment score range	Notes
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. The lexicon allocates positive or negative to words based on common use in a collection of textual data.
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. Developed by Matthew Jockers and Julia Rinker, which is more of a complex lexicon compared to Jockers, where it uses algorithms to assign a sentiment score to words based on their context and usage in a corpus.
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. Developed by Minqing Hu and Bing Liu, which assigns a positive or negative sentiment to words.
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). Assigns a sentiment score on negativity, positivity and objectivity to each synset in WordNet.
National Research Council (NRC) – filtered version	5468 words	Sentiment values of either +1 and -1.	A filtered version of NRC lexicon was developed by Mohammad & Turney's (2010) to improve accuracy and reliability, which assigns positive/negative based on the word list (RDocumentation, 2022).
Loughran Mcdonald	2702 words	Sentiment values of either +1 and -1.	Financial word list as sentiment lookup values (Loughran & Mcdonald, 2016). This lexicon is used in finance and investment.
Senticnet	23,627 words	Sentiment values ranging between -1 and 1.	Applies a combination human-generated data and natural language processing techniques to designate sentiment scores to words/ phrases. 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.	A lexicon that allocates sentiment scores for various political, rhetorical dimensions and psychological to each word. 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. A lexicon contains a collection of words and associated sentiment to classify text.
Valence Aware Dictionary and Sentiment Reasoner (VADAR)	7236 words	Sentiment values ranging between -1 and 1.	The lexicon accounts for context and the intensity of words in sentence(s), and the valence of words in relation to each other. Dataset contains a filtered version of Hutto & Gilbert's (2014) positive/negative word list as sentiment lookup values that are attuned to sentiments in social media.
Syuzhet	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 refers to the plot structure and development of a literary work. This lexicon created from 165,000 human coded terms from corpus of contemporary novels.
Bing	6789	Sentiment values (+1, -1)	The lexicon was developed by Bing Liu which is based on the emotional intensity or valence of words. This detects positive/ negative words based on sentiment scores (Liu & Hu, 2021).
AFINN	2,477 words	Ranging between - 5 (very negative) and 5 (very positive).	A word-based lexicon that designates a sentiment score to individual words/ phrases on their emotional content. The words are based on Affective Norms for English Words (Nielsen, 2011).
NRC (NRC Word-Emotion Association Lexicon)	14,182 words	sentiments: negative, positive emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, trust	Based on Mohammad & Turney (2010) paper called ”Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon.”
Sentiment Berkeley	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	Approximately 10000	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 package stansent wraps Stanford's coreNLP sentiment tagger. Tag sentiment as most negative (-1) to most positive (+1) (Rinker, 2017). The lexicon is based on Stanford Sentiment Treebank that includes sentiment annotations from movie reviews.
SentiStrength	2546	Ranging between - 5 (very negative) and 5 (very positive).	SentiStrength is a tool that is constructed by combining General Inquirer (GI) and Linguistic Inquiry and Word Count (LIWC) dictionaries and includes lists of negations, intensifiers and emoticons (Islam & Zibran, 2017; Thelwall, 2019). This assigns sentiment scores to words/ phrases based on their present in a dataset of human-generated text.

2.4 Review of How Lexicon Approach Used

The different lexicons for analysis outlined may be applied to different subject areas, such as politics, business and public [2]. 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 [2]. In the context of sentiment analysis, businesses monitor their brand reputation, competitive research and online advertising [2]. 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 [2]. 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 [2]. 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 [2]. 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) [2]. 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 [2] and 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.

As the focus of this paper is on Twitter datasets based on public order events, it would be prudent to examine existing research in this area. As cited above, 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 [2] 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 [2] into bullying by using text classification to identify various emotions, such as empathy, sadness, pride and anger in tweets. In another project [2], 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, which is limited and requires greater research, such as the dynamics of institutional application, interactions between data analysis and human intervention [2].

There has previously been some research using Twitter data based on demonstrations, such as [6] 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 [2]. Therefore, these authors decided to translate the language of the sentiment lexicon while making an application of string similarity functions. The authors used SentiWordNet as a baseline and manually created a sentiment lexicon of 6300 words to align to the context of demonstrations. The focus of [6] 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 in Birmingham, UK 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, UK, which had a peaceful event [6]. This 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 [6]. Even though there are many technical challenges to overcome, researchers, businesses and organisations continue to strive for new techniques (or to combine existing methods) to achieve higher levels of accuracy and representativeness in sentiment analysis [2].

2.5 Adapting Lexicons for Sentence Level

We are comparing a range of sentiment analysis methods based on sentence level, we will receive Twitter data as input and produce a both polarity score and category as an output [2]. As noted in Table 1 there are different ranges for the 18 lexicon-based dictionaries to classify the relevant tweets from -1 to +1 to -5 to +5, but most are in the range of -1 to +1, therefore, we standardised the range for all dictionaries results to conform to this range. 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 [2]. Therefore, the difference in the outcome is not significant when the scores a rescaled in the same range.

As previously stated above, the output for each method can vary depending on its what it was developed for and its approach. Furthermore, some dictionaries listed weight of terms can be rather small which may impact the results [2]. Therefore, based on these points, we created 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 sentiment scores. The majority of 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 of Hiu Liu, Jockers, Jockers Rinker, Loughran MacDonald, NRC, SenticNet4, SentiWordNet, Slangsd, Inquirer, Vadar and AFINN lexicon-based dictionaries [2]. These dictionaries have been selected on the basis that if the scored word list is similar, larger difference of words to expand the list and whether the list is available to extract the terms with their weightings. 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 results [2].

Some of the initial sentiment analysis results on the tweets showed promising F1 scores of 0.60s for many of the dictionaries, but the combined dictionary showed no sentiment results for neutral category as there are no scores on 0 [2]. The tweets near score of 0 on closer inspection show many tweets that should have been classified as neutral. We decided to implement a cut-off threshold to classify tweets that should be neutral, 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 F1 score of each one’s result. The one with the more evenly balanced precision, recall and F1 for each sentiment category on each dataset will be chosen as the cut-off point [2].

The manually coded (relevant data) results with no threshold tended to be lowest F1 score except for Dover being its highest F1 score, which is understandable given the data is mainly negative in sentiment [2]. Additionally, the no threshold results for neutral are of 0 recall and precision 1 as no neutral results exists. The highest F1 score for all datasets tended to be 1, but the unevenness between the precision and recall is high. The higher F1 score reduces precision for negative and recall increases, putting this out of balance across the sentiment categories. Thus, a lower F1 score with a cut-off of 0.5 produces the best performance with a more evenly spread precision and recall across the sentiment categories for each dataset. Therefore, the cut-off range of 0.5 to -0.5 is chosen for the combined dictionary to classify tweets as neutral [2]. The automated coded (relevant data) results from the combined dictionary contains no neutrals similar to the manually coded (relevant data). The same process for the cut-off was repeated for the automatically selected relevant data and 0.5 cut-off appeared an unreasonable choice due to low F1 scores because of imbalance between the sentiment categories. There was an incline in F1 scores within both 0.5 and 1, but the highest was between the ranges of 0.6 to 1. The highest F1 for both 2016 MMM and 2016 Anti-Austerity is 1 except for 2016 Dover which peaked at 0.3 and 2015 MMM at 0.9 [2]. The spread of precision, recall and f-measure is reasonably balanced at 1 for most of the dataset’s results. We compared both manual and automated, and decided that a cut-off of 0.5 was applied as the manual coded (relevant data) was more unevenly balanced with lower F1 scores, and also automated range was near 0.5 with best scores ranging from 0.6 to 1 with lesser difference between nearing to 1 in terms of F1 score [2].

3 Applied Datasets and Lexicon Approach

The first step of the hybrid approach is data extraction which has already been collected from Twitter which are based on demonstrations that took place both 2015 and 2016 Million Mask March (MMM), 2016 Anti-Austerity March, and 2016 Dover events [2].

The second step is coding the data based on relevant and irrelevant data for each event. 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 [2]. A small sample of tweets will be manually coded for each dataset. The keywords listed built will be used to automatically code (relevant) tweets from the dataset for each event. The relevant tweets have been both manually and automatically coded. 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 2 used for each dataset [2]. This total number does not include retweets which are automatically removed from the datasets in the cleansing phase. Table 2 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 not identified as relevant which may due to the lack of keywords used in the list. As a result, the proportion of relevant and irrelevant tweets was lower than expected. Consequently, the keywords lists were extended with new words to increase relevant and irrelevant categories [2].

Table 2 Classification of relevance results

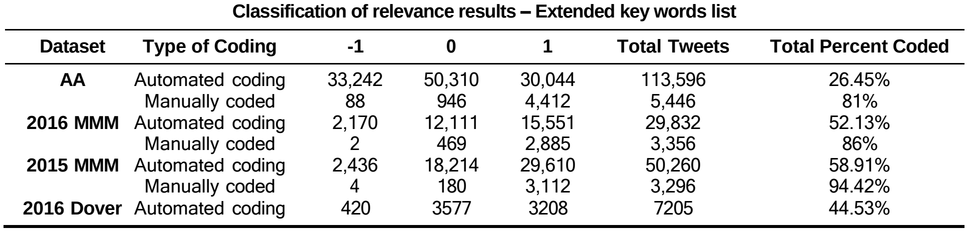
Classification of relevance results

Dataset	Type of Coding	-1	0	1	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%

The keywords list was extended by adding both the most frequently counted words and Frequency–Inverse Document Frequency (TF-IDF) [2]. As a result, of this change the number of relevant and irrelevant tweets increased with fewer being unclassified. Table 3 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% [2].

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 [2]. Both manual and automated coded (relevant) tweets are pre-processed

Table 3 Classification of relevance results – Extended key words list



Classification of relevance results – Extended key words list

Dataset	Type of Coding	-1	0	1	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%

with data cleansing techniques applied. The automated relevant data will be used for sentiment analysis process with the 19 dictionaries including the combined dictionary. The cleansed tweets for each dataset will be validated for its reliability. 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 pre-processing of the data has left some tweets blank with no score which are removed from each dataset, so there are now Anti-Austerity 29,963, 2016 Dover 3,174, 2016 MMM 15,491 and 2015 MMM 29,420 [2].

The 19 dictionaries are applied to both the manual coded (relevant tweets) and the automated coded data (relevant tweets) [2]. The standardisation process (ensuring in a range of -1 to +1 as some dictionaries varied, much easier to compare if within set range) was applied to specific lexicon dictionaries as previously described. An important aspect of evaluating the sentiment analysis approach is the use of accurate gold standard labelled datasets of which several already exist produced by expert and non-expert human annotators. For this study, we will use the four Twitter datasets in which we attempt to assess the quality of our gold standard datasets in terms of the accuracy of the labelling process. A 1500 tweets of each dataset (based on manually coded relevant data) will be evaluated to validate the dictionaries' reliability [2].

In Table 4, each tweet in the sample is manually classified by a series of non-expert users that are known as Manual Rater 1 (MR1) and Manual Rater 2 (MR2) to measure the reliability [2]. Table 4 results from the inter-agreement have shown MR1 and MR2 to be reliable, as shows a high level of agreement for 3 of the datasets with 2015 MMM 70.3% agreement, 2016 MMM 71.3%, 2016 Dover 76.3%, but 2016 Anti-Austerity show much less agreement between MR1 and MR2 on 56%. Furthermore, Table 4 shows 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 [2]. 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. This reasonably high level of agreement has shown that 'Gold Standard' can be used as a baseline against new data in the hybrid approach. 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 [2]. The Gold Standard will be evaluated against the sentiment analysis results, which include the analysis techniques of precision, recall, F1 score and proportion that agreed between each sentiment category. These evaluation techniques are defined in the next section.

Table 4 Inter Agreement Results

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	24.4	62.8	12.33
Neutral	34.73	40	12.67	38.73
Positive	3.6	5.87	0.87	4.93
Disagree	28.73	29.73	23.67	44
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

3.1 Evaluation Techniques

We will present the comparison of results for 19 dictionaries on the four UK demonstration event datasets. The dictionaries will be applied to identify which tweets are positive, negative and neutral. The comparison of the 3-class comparison, we used Precision, Recall and F1 measures for classification to determine the accuracy of each lexicon results. 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 [2].

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

Recall determines the number of elements that have been retrieved over the total number of relevant instances [2]. This is defined as the number of true positives over the total number of positives.

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

Precision and recall can be combined into F1 or F-Measure, which measures the accuracy of the classification as a whole [2]. F1 takes account of both precision and recall [2]. 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:

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

F-measure can describe the model's performance with a singular number enabling comparisons across several models against one another [2]. F1 can apply different weights to calculate the F-score for precision and recall, but it may be difficult to assign appropriate weights [2]. 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 [2].

F-measure is useful to measure the performance of text classification in a way that is informative and more useful than classification accuracy [2]. 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 [2]. Multi-class text classification performance is measured on the effectiveness based on macro-averaged and micro-averaged of F-measure scores [2]. 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 [2]. 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 [2]. These indicators should not be a way to determine how reliable a classifier will be for future performance on unseen data [2]. 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 [2].

The evaluation techniques explored for sentiment analysis can help understand how conclusive the results of any sentiment classification result. The proceeding section 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 for both MR1 and MR2.

4 Sentiment Analysis Results

We will begin the analysis of our experiment comparing the results for all datasets based on the 3-class comparison for all dictionaries, which will present Table 6 with the precision, recall, F1 score, Micro-F1, and Macro-F1. A final comparison will be made with our results compared with other research papers, for example, comparing [8] three-class based results on 24 dictionaries, but these authors did explore two-class based as well. We considered a five-class approach as some tweets were not fully negative, but rather somewhat negative, so this is where a five-class category was created but the F1 scores were poor on the initial results, it appeared more class categories increased mis-classification rate, as more difficult to assign with more categories, therefore, five-class approach was dropped, which is why not included in the series of results.

In Table 5, 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. MR1 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. 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 5 and Table 6 where Macro average is lower than the Micro average, there is poor metric performance on the smaller classes [2]. 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.

Table 5 shows 2015 MMM Jockers family with 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.

Table 5 MR1 3-classes experiments results with 4 datasets

Dataset	Method	Positive Sentiment			Negative Sentiment			Neutral Sentiment			Macro-F1	Micro-F1
		P	R	F1	P	R	F1	P	R	F1		
2015 MMM	combined_dictionary	0.42	0.11	0.18	0.51	0.61	0.56	0.38	0.67	0.49	0.45	0.47
	berkeley	0.7	0.08	0.14	0.54	0.53	0.53	0.05	0.52	0.08	0.4	0.26
	inquirer	0.47	0.18	0.26	0.58	0.48	0.52	0.65	0.71	0.68	0.51	0.58
	jockers_rinker	0.6	0.17	0.27	0.54	0.81	0.65	0.37	0.9	0.52	0.56	0.54
	loughran_mcdonald	0.13	0.19	0.16	0.51	0.52	0.52	0.68	0.66	0.67	0.45	0.59
	nrc	0.57	0.1	0.17	0.54	0.26	0.35	0.52	0.68	0.59	0.42	0.43
	senticnet	0.69	0.09	0.16	0.48	0.62	0.54	0.04	0.71	0.08	0.44	0.29
	sentistrength	0.14	0.08	0.1	0.38	0.44	0.41	0.48	0.58	0.53	0.35	0.44
	slangsd	0.16	0.07	0.1	0.37	0.41	0.39	0.5	0.64	0.56	0.36	0.45
	socal_google	0.32	0.07	0.12	0.51	0.21	0.3	0.6	0.63	0.62	0.37	0.44
	stanford	0.1	0.09	0.1	0.39	0.45	0.42	0.53	0.58	0.55	0.36	0.47
	vadar	0.65	0.2	0.3	0.57	0.66	0.61	0.52	0.81	0.64	0.57	0.58
	sentimentr_huliu	0.47	0.22	0.3	0.58	0.56	0.57	0.65	0.73	0.69	0.53	0.61
	sentimentr_jockers	0.61	0.17	0.27	0.55	0.8	0.65	0.39	0.89	0.54	0.56	0.55
	sentimentr_sentiword	0.48	0.08	0.13	0.47	0.6	0.53	0.18	0.77	0.29	0.42	0.35
	syuzhet_afinn	0.49	0.2	0.28	0.57	0.67	0.62	0.56	0.78	0.66	0.55	0.6
	syuzhet_bing	0.48	0.22	0.31	0.58	0.53	0.55	0.66	0.72	0.69	0.53	0.6
	syuzhet_jockers	0.62	0.18	0.28	0.55	0.81	0.65	0.39	0.9	0.54	0.57	0.55
	syuzhet_nrc	0.58	0.1	0.17	0.54	0.25	0.34	0.53	0.66	0.59	0.42	0.43
2016 MMM	combined_dictionary	0.53	0.21	0.3	0.44	0.59	0.5	0.37	0.68	0.48	0.47	0.45
	berkeley	0.69	0.13	0.22	0.41	0.4	0.4	0.1	0.51	0.17	0.37	0.25
	inquirer	0.54	0.32	0.4	0.45	0.37	0.4	0.68	0.72	0.7	0.51	0.58
	jockers_rinker	0.72	0.28	0.4	0.46	0.71	0.56	0.39	0.89	0.54	0.57	0.52
	loughran_mcdonald	0.24	0.37	0.29	0.44	0.46	0.45	0.73	0.7	0.71	0.49	0.6
	nrc	0.6	0.17	0.27	0.42	0.27	0.33	0.49	0.72	0.58	0.44	0.44
	senticnet	0.78	0.16	0.27	0.37	0.54	0.44	0.05	0.8	0.1	0.45	0.27
	sentistrength	0.43	0.38	0.4	0.44	0.55	0.49	0.62	0.72	0.67	0.52	0.58
	slangsd	0.11	0.12	0.12	0.3	0.48	0.37	0.49	0.67	0.56	0.35	0.44
	socal_google	0.53	0.19	0.28	0.38	0.16	0.23	0.61	0.69	0.65	0.41	0.48
	stanford	0.16	0.18	0.17	0.32	0.45	0.37	0.54	0.64	0.59	0.38	0.47
	vadar	0.72	0.3	0.42	0.5	0.6	0.55	0.52	0.83	0.64	0.58	0.57
	sentimentr_huliu	0.61	0.35	0.44	0.46	0.44	0.45	0.64	0.74	0.69	0.54	0.58
	sentimentr_jockers	0.72	0.27	0.39	0.46	0.69	0.56	0.4	0.88	0.55	0.57	0.52
	sentimentr_sentiword	0.69	0.17	0.27	0.33	0.48	0.39	0.13	0.67	0.22	0.41	0.29
	syuzhet_afinn	0.58	0.3	0.39	0.49	0.6	0.54	0.56	0.79	0.66	0.55	0.58
	syuzhet_bing	0.58	0.34	0.43	0.47	0.44	0.46	0.65	0.73	0.68	0.53	0.58
	syuzhet_jockers	0.7	0.27	0.39	0.46	0.7	0.55	0.39	0.87	0.54	0.56	0.51
	syuzhet_nrc	0.6	0.18	0.27	0.42	0.26	0.32	0.5	0.71	0.58	0.44	0.45
2016 AA	combined_dictionary	0.69	0.18	0.28	0.24	0.54	0.34	0.34	0.85	0.49	0.47	0.4
	berkeley	0.83	0.13	0.23	0.19	0.45	0.27	0.04	0.72	0.08	0.39	0.18
	inquirer	0.6	0.22	0.32	0.2	0.31	0.24	0.56	0.83	0.67	0.45	0.53
	jockers_rinker	0.76	0.2	0.32	0.23	0.59	0.33	0.31	0.91	0.46	0.49	0.39
	loughran_mcdonald	0.27	0.3	0.29	0.19	0.3	0.23	0.73	0.8	0.77	0.43	0.63
	nrc	0.63	0.19	0.29	0.25	0.44	0.32	0.47	0.83	0.6	0.47	0.48
	senticnet	0.81	0.13	0.23	0.18	0.44	0.25	0.04	0.93	0.07	0.41	0.17
	sentistrength	0.49	0.3	0.37	0.24	0.59	0.34	0.55	0.84	0.66	0.49	0.55
	slangsd	0.06	0.06	0.06	0.14	0.44	0.21	0.46	0.77	0.58	0.29	0.42
	socal_google	0.45	0.12	0.19	0.18	0.14	0.15	0.56	0.82	0.66	0.38	0.49
	stanford	0.4	0.3	0.34	0.22	0.65	0.33	0.51	0.84	0.63	0.46	0.52
	vadar	0.74	0.2	0.32	0.24	0.48	0.32	0.4	0.88	0.55	0.49	0.45
	sentimentr_huliu	0.63	0.24	0.35	0.25	0.47	0.33	0.55	0.86	0.67	0.5	0.54
	sentimentr_jockers	0.77	0.2	0.32	0.23	0.58	0.33	0.32	0.91	0.47	0.49	0.4
	sentimentr_sentiword	0.69	0.13	0.22	0.16	0.42	0.23	0.11	0.86	0.2	0.38	0.21
	syuzhet_afinn	0.73	0.24	0.36	0.23	0.47	0.3	0.45	0.86	0.59	0.49	0.48
	syuzhet_bing	0.6	0.24	0.35	0.26	0.47	0.34	0.57	0.85	0.68	0.5	0.56
	syuzhet_jockers	0.76	0.2	0.31	0.24	0.61	0.35	0.32	0.91	0.47	0.5	0.4
	syuzhet_nrc	0.63	0.19	0.29	0.26	0.44	0.32	0.47	0.83	0.6	0.47	0.48
2016 Dover	combined_dictionary	0.61	0.09	0.15	0.76	0.59	0.66	0.31	0.37	0.34	0.43	0.5
	berkeley	0.78	0.06	0.11	0.74	0.56	0.64	0.07	0.29	0.11	0.39	0.41
	inquirer	0.57	0.07	0.13	0.79	0.4	0.54	0.55	0.42	0.48	0.41	0.46
	jockers_rinker	0.76	0.09	0.16	0.79	0.72	0.76	0.21	0.56	0.31	0.51	0.56
	loughran_mcdonald	0.29	0.13	0.18	0.74	0.38	0.5	0.69	0.37	0.48	0.39	0.48
	nrc	0.53	0.07	0.13	0.8	0.45	0.58	0.51	0.41	0.46	0.41	0.47
	senticnet	0.8	0.05	0.09	0.72	0.47	0.57	0.04	0.42	0.07	0.39	0.34
	sentistrength	0.53	0.12	0.2	0.79	0.69	0.74	0.43	0.48	0.45	0.49	0.6
	slangsd	0.06	0.01	0.02	0.67	0.43	0.52	0.49	0.35	0.41	0.32	0.44
	socal_google	0.37	0.03	0.06	0.77	0.25	0.38	0.47	0.34	0.4	0.3	0.33
	stanford	0.27	0.08	0.12	0.67	0.47	0.55	0.41	0.31	0.35	0.35	0.45
	vadar	0.69	0.09	0.15	0.79	0.67	0.73	0.31	0.53	0.39	0.5	0.56
	sentimentr_huliu	0.59	0.09	0.15	0.79	0.56	0.66	0.41	0.42	0.42	0.45	0.52
	sentimentr_jockers	0.78	0.09	0.16	0.79	0.71	0.75	0.23	0.55	0.32	0.51	0.56
	sentimentr_sentiword	0.67	0.04	0.08	0.72	0.45	0.55	0.12	0.39	0.18	0.37	0.35
	syuzhet_afinn	0.61	0.1	0.17	0.78	0.67	0.72	0.38	0.48	0.42	0.49	0.57
	syuzhet_bing	0.55	0.08	0.14	0.79	0.52	0.63	0.45	0.4	0.43	0.43	0.5
	syuzhet_jockers	0.73	0.08	0.15	0.78	0.71	0.74	0.21	0.55	0.31	0.5	0.55
	syuzhet_nrc	0.49	0.07	0.12	0.8	0.43	0.56	0.51	0.4	0.45	0.4	0.46

In Table 5, 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.

In Table 6, 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. 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.

Table 6 MR2 3-classes experiments results with 4 datasets

Dataset	Method	Positive Sentiment			Negative Sentiment			Neutral Sentiment			Macro-F1	Micro-F1
		P	R	F1	P	R	F1	P	R	F1		
2015 MMM	combined dictionary	0.57	0.13	0.21	0.77	0.59	0.67	0.45	0.52	0.48	0.49	0.53
	berkeley	0.8	0.08	0.14	0.8	0.5	0.61	0.04	0.29	0.07	0.38	0.34
	inquirer	0.69	0.23	0.34	0.87	0.46	0.6	0.78	0.56	0.65	0.54	0.59
	jockers_rinker	0.81	0.21	0.33	0.82	0.78	0.8	0.52	0.84	0.64	0.66	0.68
	loughran_mcdonald	0.22	0.28	0.25	0.8	0.52	0.63	0.82	0.53	0.65	0.52	0.62
	nrc	0.65	0.1	0.17	0.84	0.26	0.4	0.64	0.55	0.59	0.43	0.43
	senticnet	0.81	0.09	0.16	0.72	0.59	0.65	0.05	0.54	0.09	0.46	0.39
	sentistrength	0.17	0.08	0.11	0.56	0.41	0.47	0.46	0.38	0.41	0.33	0.41
	slangsd	0.13	0.05	0.07	0.59	0.42	0.49	0.53	0.46	0.49	0.36	0.45
	socal_google	0.33	0.06	0.1	0.74	0.2	0.31	0.67	0.47	0.55	0.34	0.39
	stanford	0.1	0.08	0.09	0.61	0.45	0.52	0.57	0.42	0.49	0.37	0.48
	vadar	0.84	0.22	0.35	0.84	0.62	0.71	0.67	0.7	0.68	0.62	0.65
	sentimentr_huliu	0.65	0.26	0.38	0.86	0.54	0.66	0.78	0.59	0.67	0.58	0.64
	sentimentr_jockers	0.83	0.21	0.33	0.82	0.77	0.79	0.53	0.82	0.65	0.66	0.68
	sentimentr_sentiword	0.58	0.08	0.14	0.7	0.57	0.63	0.22	0.65	0.33	0.47	0.44
	syuzhet_afinn	0.7	0.24	0.36	0.82	0.62	0.71	0.69	0.64	0.66	0.6	0.65
	syuzhet_bing	0.67	0.27	0.39	0.86	0.5	0.63	0.79	0.57	0.66	0.57	0.62
	syuzhet_jockers	0.81	0.2	0.33	0.82	0.77	0.8	0.53	0.83	0.65	0.66	0.68
	syuzhet_nrc	0.65	0.1	0.17	0.87	0.26	0.4	0.64	0.53	0.58	0.42	0.43
2016 MMM	combined dictionary	0.62	0.18	0.27	0.69	0.55	0.61	0.4	0.55	0.47	0.49	0.49
	berkeley	0.74	0.1	0.18	0.72	0.42	0.53	0.08	0.33	0.13	0.36	0.29
	inquirer	0.71	0.3	0.42	0.75	0.37	0.49	0.75	0.6	0.67	0.54	0.57
	jockers_rinker	0.86	0.23	0.37	0.72	0.67	0.69	0.48	0.83	0.6	0.63	0.6
	loughran_mcdonald	0.35	0.39	0.37	0.72	0.45	0.55	0.81	0.59	0.68	0.54	0.61
	nrc	0.74	0.15	0.25	0.72	0.28	0.4	0.55	0.62	0.59	0.46	0.44
	senticnet	0.8	0.12	0.21	0.61	0.52	0.56	0.06	0.72	0.12	0.47	0.33
	sentistrength	0.65	0.4	0.5	0.74	0.55	0.63	0.74	0.65	0.69	0.61	0.64
	slangsd	0.09	0.07	0.07	0.49	0.48	0.49	0.52	0.53	0.52	0.36	0.46
	socal_google	0.65	0.16	0.26	0.7	0.18	0.28	0.64	0.54	0.59	0.41	0.43
	stanford	0.21	0.17	0.19	0.53	0.45	0.49	0.57	0.52	0.54	0.41	0.49
	vadar	0.87	0.25	0.39	0.8	0.57	0.66	0.62	0.75	0.68	0.62	0.62
	sentimentr_huliu	0.74	0.3	0.43	0.76	0.44	0.55	0.73	0.63	0.67	0.56	0.6
	sentimentr_jockers	0.85	0.23	0.36	0.73	0.65	0.69	0.49	0.81	0.61	0.62	0.59
	sentimentr_sentiword	0.73	0.13	0.22	0.56	0.48	0.52	0.15	0.58	0.24	0.43	0.35
	syuzhet_afinn	0.77	0.28	0.41	0.77	0.56	0.65	0.66	0.7	0.68	0.61	0.63
	syuzhet_bing	0.73	0.31	0.43	0.78	0.43	0.56	0.73	0.62	0.67	0.56	0.6
	syuzhet_jockers	0.83	0.23	0.36	0.73	0.67	0.7	0.48	0.8	0.6	0.62	0.6
	syuzhet_nrc	0.71	0.15	0.25	0.71	0.27	0.39	0.55	0.6	0.57	0.45	0.44
2016 AA	combined dictionary	0.83	0.19	0.31	0.73	0.45	0.55	0.4	0.57	0.47	0.5	0.46
	berkeley	0.91	0.13	0.23	0.69	0.45	0.54	0.03	0.31	0.06	0.38	0.31
	inquirer	0.78	0.25	0.38	0.76	0.32	0.45	0.71	0.59	0.65	0.51	0.53
	jockers_rinker	0.93	0.22	0.35	0.79	0.57	0.66	0.44	0.74	0.55	0.6	0.55
	loughran_mcdonald	0.37	0.37	0.37	0.78	0.34	0.48	0.86	0.53	0.66	0.51	0.57
	nrc	0.73	0.2	0.31	0.82	0.39	0.53	0.56	0.57	0.56	0.5	0.5
	senticnet	0.89	0.13	0.22	0.61	0.42	0.5	0.05	0.74	0.1	0.47	0.31
	sentistrength	0.78	0.42	0.55	0.81	0.57	0.67	0.74	0.64	0.69	0.64	0.66
	slangsd	0.09	0.07	0.08	0.52	0.47	0.49	0.51	0.49	0.5	0.36	0.45
	socal_google	0.55	0.13	0.22	0.77	0.16	0.27	0.63	0.52	0.57	0.38	0.4
	stanford	0.52	0.34	0.41	0.74	0.61	0.67	0.65	0.6	0.63	0.57	0.62
	vadar	0.91	0.22	0.35	0.81	0.45	0.58	0.55	0.68	0.61	0.56	0.54
	sentimentr_huliu	0.83	0.28	0.42	0.83	0.43	0.57	0.7	0.62	0.66	0.57	0.58
	sentimentr_jockers	0.94	0.22	0.35	0.81	0.57	0.67	0.45	0.73	0.56	0.6	0.55
	sentimentr_sentiword	0.78	0.13	0.22	0.64	0.46	0.53	0.14	0.61	0.23	0.45	0.35
	syuzhet_afinn	0.88	0.25	0.39	0.79	0.45	0.58	0.6	0.64	0.61	0.56	0.55
	syuzhet_bing	0.79	0.29	0.42	0.84	0.41	0.55	0.71	0.6	0.65	0.56	0.58
	syuzhet_jockers	0.95	0.22	0.36	0.81	0.57	0.67	0.45	0.73	0.56	0.6	0.55
	syuzhet_nrc	0.75	0.2	0.32	0.82	0.39	0.53	0.56	0.55	0.56	0.5	0.49
2016 Dover	combined dictionary	0.86	0.07	0.13	0.92	0.57	0.7	0.41	0.25	0.31	0.42	0.55
	berkeley	0.83	0.04	0.07	0.92	0.55	0.69	0.08	0.17	0.1	0.36	0.48
	inquirer	0.72	0.06	0.1	0.95	0.38	0.55	0.66	0.25	0.37	0.36	0.43
	jockers_rinker	0.86	0.06	0.11	0.95	0.69	0.8	0.34	0.45	0.39	0.51	0.64
	loughran_mcdonald	0.34	0.09	0.15	0.94	0.38	0.54	0.8	0.22	0.34	0.35	0.45
	nrc	0.55	0.04	0.08	0.97	0.43	0.59	0.62	0.25	0.36	0.36	0.46
	senticnet	0.9	0.03	0.06	0.89	0.46	0.6	0.07	0.36	0.11	0.39	0.4
	sentistrength	0.79	0.11	0.19	0.96	0.66	0.78	0.62	0.34	0.44	0.5	0.66
	slangsd	0	0	0	0.88	0.44	0.59	0.6	0.22	0.32	0.3	0.46
	socal_google	0.69	0.04	0.07	0.94	0.24	0.39	0.59	0.21	0.31	0.27	0.31
	stanford	0.38	0.06	0.11	0.85	0.48	0.61	0.46	0.17	0.25	0.34	0.47
	vadar	0.86	0.06	0.12	0.95	0.64	0.77	0.46	0.4	0.43	0.5	0.62
	sentimentr_huliu	0.83	0.07	0.13	0.96	0.54	0.69	0.56	0.28	0.38	0.43	0.55
	sentimentr_jockers	0.9	0.06	0.11	0.95	0.68	0.79	0.38	0.46	0.41	0.52	0.63
	sentimentr_sentiword	0.79	0.03	0.06	0.89	0.44	0.59	0.18	0.29	0.22	0.36	0.4
	syuzhet_afinn	0.69	0.07	0.12	0.94	0.64	0.76	0.51	0.33	0.4	0.46	0.62
	syuzhet_bing	0.97	0.09	0.16	0.96	0.5	0.66	0.61	0.27	0.38	0.43	0.53
	syuzhet_jockers	0.9	0.06	0.11	0.95	0.68	0.79	0.36	0.46	0.41	0.52	0.64
	syuzhet_nrc	0.55	0.04	0.08	0.97	0.41	0.58	0.61	0.24	0.34	0.35	0.45

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. 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 F-measure 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 F1 scores are for macro/micro 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 precision, recall, F1 and macro F1 have been explored for each dictionary to determine the strength of the results of which most macro F1 are scored in 0.60s for 3-class experiment, and this shows there is improvement to made within the 19 lexicon dictionaries. As previously noted, the best method varies from one dataset to another, however, Jockers family (Jockers Rinker, Syuzhet Jockers, Sentimentr Jockers are similar with a degree of variation) tend to appear most consistently in best method as indicated by the macro F1 scores, but Vadar, Syuzhet Bing and SentiStrength have excelled albeit not best method for some datasets within both MR1 and MR2 results, but respectively perform well as appear in top 10 or mid-way out of the 19 dictionaries. However, when compared to mean rank value based on macro-F1 averaged across each dataset for MR1 are ranked Sentimentr Jockers, Syuzhet Jockers, Jockers rinker, Vadar, Syuzhet AFINN, Sentimentr Huliu, Syuzhet Bing, SentiStrength and Inquirer for the first several dictionaries, which is near similar to MR2 results, but there are some differences, such as Vadar ranked 1 for MR2 rather than ranked 3 in MR1 and Combined Dictionary being above Inquirer in the ranking by one place.

Most methods are better to classify negative and neutral sentences than positive, but some dictionaries appear to identify more positives, such as SentiStrength and Vadar out of the 19 dictionaries, but still negative, and neutral remain higher in terms of numbers classified, so this suggests some dictionaries are somewhat more bias to positivity [2]. The bias observed in the sentiment analysis results are based on the way the dictionary was designed (e.g., aligned towards a topic in building a lexicon, words contained, and how assigned weightings) and the context of the event, as some events can be more positive or negative depending on the topic. This impacts the sentiment analysis results which is why there is a degree of variation when evaluating the output, where some dictionaries are stronger and other ones produce less strong results dependant on the context of the topic [2]. Also, 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. The manual classification of tweets classified needs to be increased for greater generalisation. 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. 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 [2].

In the next section, we compare both MR1 and MR2 results with other research papers results to identify if there are similarities and/ or differences in the outcomes.

4.1 Comparison With Other Published Results

In review of existing research, we discovered that a range of different papers focused on either 2-classes, 3-classes, or both in their lexicon-based approach. However, as cited above (refer to section 2.2) there is limited research on comparing lexicons, therefore, this limits comparison with our research and a further difficulty is papers refer to accuracy rather than precision, recall, and F1 scores which can make it challenging to make comparisons with our research. We will compare similar precision, recall and F1 scores results of 3 classes used in a lexicon-based approach.

Methods tended to perform better in the context they were originally evaluated [8] which is expected, and also depends on how well tuned the process was on the acquisition of data to applying sentiment analysis methods. Now that we have compared MR1 and MR2 mean rank, it would be good to compare with [8] which known as there “SentiBench” approach, part of the results “calculated the mean rank for these methods without their ‘original’ datasets and put the results in parenthesis” for social network data which is shown in Table 7. Also, [8] emphasises could only compare SentiStrength and VADER due to “kindly allowed the entire reproducibility of their work, sharing both methods and datasets.” It is important to note that some of the dictionaries are applied, therefore, some comparison can be drawn. For instance, in Table 7 the SentiBench results shows Vadar and AFINN are in the top 5 similarly to both MR1 and MR2, and also SentiStrength appearing in top 10, and furthermore, SentiWordNet ranked 13 in a similar position to MR1/ MR2 on 12th, but mean rank is near 13.57. According to [5] VADAR is attuned to sentiments in social media, which in turn could be the reason it shows a strong consistent performance in both our and SentiBench results with other research shows (1,2 citations below). Furthermore, both SenticNet and WordNet appear widely used (which SentiWordNet assigns scores to WordNet sysnets) in Lexicon based approach for different contexts [5]. SentiBench [8] shows Senticnet on 14th, whereas both MR1 and MR2 are 11th position performing slightly better on the Twitter dataset for these set events. The other dictionaries that are not comparable with SentiBench, such as Bing, Hiu Liu, AFINN and Jockers are widely used in research that tend to perform well in a range of contexts [7] which appear in the top 10 for both MR1 and MR2, which has shown Hu Liu performed highly on product review data on 0.76 accuracy, but the Inqurier, NRC and SentiWordNet were dropped in their experimental study as produced weaker results. In our study shows them one of the better performers, overall, this goes to show that there is variation depending on the dataset it is applied too, but there are consistently good performers in the top 10 that have worked with the Twitter data, and similarly in SentiBench datasets, which shows generally good across a wide range of contexts, whereas some others perform better in specific contexts that they were designed for.

As we noted along with SentiBench [8], it is important that a standard sentiment analysis benchmark is needed and constantly updated which we understand and have looked to do in this paper. Furthermore, there are other open available methods we may have not evaluated, but more specifically paid options as cited in [8] could use more gold standard datasets and increase the variety of datasets from social media platforms to enhance their lexicon approach to improve the accuracy of results on various different topics. The creation of the combined dictionary was to identify if the inclusion of other lexicons thus making a very large pool of words weighted would enhance the output, but as the results show it appeared mid-way or 9th position out of the 19 dictionaries, thus the inclusion of more words does not necessarily mean it will be top or near the top in ranking out of the dictionaries.

Table 7 Compare mean rank for datasets

MR1 3-Classes			**MR2 3-Classes**		
Method	**Mean Rank**	**Position**	**Method**	**Mean Rank**	**Position**
sentimentr_jockers	1.5	1	vadar	2.25	1
syuzhet_jockers	1.5	1	syuzhet_jockers	2.25	1
jockers_rinker	1.75	2	jockers_rinker	2.5	2
vadar	4.25	3	sentimentr_jockers	2.5	2
syuzhet_afinn	5.75	4	syuzhet_afinn	4.75	3
sentimentr_huliu	6.25	5	sentimentr_huliu	5	4
syuzhet_bing	7	6	syuzhet_bing	5.5	5
sentistrength	7.25	7	sentistrength	9	6
inquirer	9.5	8	combined dictionary	9.25	7
combined_dictionary	10.5	9	inquirer	10	8
loughran_mcdonald	10.75	10	nrc	11	9
nrc	12.25	11	loughran_mcdonald	11.5	10
senticnet	12.25	11	syuzhet_nrc	11.5	10
sentimentr_sentiword	13.25	12	senticnet	12.75	11
stanford	13.5	13	sentimentr_sentiword	15	12
syuzhet_nrc	13.75	14	berkeley	15.5	13
berkeley	15.25	15	stanford	15.75	14
socal_google	17.5	16	socal_google	16.75	15
slangsd	18	17	slangsd	18.25	16
SentiBench 3-Classes (Social Networks)					
Method	**Mean Rank**	**Position**			
Umigon	2.57	1			
LIWC15	3.29	2			
VADER	4.57(4.57)	3			
AFINN	5	4			
Opinion Lexicon	5.57	5			
Semantria	6	6			
Sentiment140	7	7			
Pattern.en	7.57	8			
SO-CAL	9	9			
Emolex	12.29	10			
SentiStrength	12.43(11.60)	11			
Opinion Finder	13	12			
SentiWordNet	13.57	13			
SenticNet	14.14	14			
SASA	14.86	15			
LIWC	15.43	16			
Sentiment140 L	15.43	17			
USent	16	18			
ANEW SUB	19.14	19			

As mentioned above, 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. However, the 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. Furthermore, if this dictionary would be applied to a UK context, then the terms applied may need to be revised to improve the combined dictionary results 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 [2]. 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. Also, the combined dictionary would be publicly made available to help further this in some way. Also, these results may help progress the open-source dictionaries, and also the paid options to improve the method to enhance the overall output.

5 Concluding remarks

There are a wide range of sentiment analysis methods that have been developed with a range of specialists in both psychology and linguistics, which has been applied on a larger volume of data to analyse the sentiment of variety of different structured and unstructured data, such as on social media platforms to other web content, which are either short or long messages depending on the platform. We have presented a comparison of 18 sentence-level sentiment analysis methods that are widely used in various contexts along with a new combined dictionary consisting of 11 dictionaries using Twitter data based on four public order events. We have focused on the prediction performance of 18 established dictionaries and new combined dictionary to understand their impact across four Twitter datasets on three classes e.g., positive, negative, and neutral.

In the findings, we highlighted that there is some consistency shown on the best lexicon methods (refer to section 4.1, and a degree of variation as well, ones where fluctuate as depending on their design and dataset as shown in MR1, MR2 and SentiBench, alongside other papers discussed in section 4.1. This demonstrates progression as these results depict consistency of strength for some dictionaries in the results, but a larger number show more a degree of variation. Therefore, one needs to be careful on their selection of dictionary applied to the nature of their dataset, as the results show some dictionaries, such as Slangsd and Socal Google have not performed well specifically on the four-demonstration dataset from Twitter in this study.

Author details

[1]Department of Computing, Sheffield Hallam University, Sheffield, UK. [2]Dept of Statistics, University of Warwick, Coventry, UK. [3]Faculty of Science, Technology, Engineering Mathematics, School of Mathematics Statistics, The Open University, UK.[4]Department of Computing, Sheffield Hallam University, Sheffield, UK.

References

1. Brunsdon Teresa Gaudoin Jotham Baldwin, James and Laurence Hirsch. Towards a social media research methodology: Defining approaches and ethical concerns. 2018.
2. James Baldwin. *Analysing social media data using sentiment analysis in relation to public order*. PhD thesis, 2021.
3. Pollyanna Gonc¸alves, Matheus Arau´jo, Fabr´ıcio Benevenuto, and Meeyoung Cha. Comparing and combining sentiment analysis methods. Association for Computing Machinery, 2013.
4. Martin Haselmayer and Marcelo Jenny. Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. *Quality & Quantity*, 51:2623 – 2646, 2016.
5. H.S. Hota, Dinesh K. Sharma, and Nilesh Verma. 14 - lexicon-based sentiment analysis using twitter data: a case of covid-19 outbreak in india and abroad. pages 275–295, 2021.
6. Anna Jurek, Maurice D. Mulvenna, and Yaxin Bi. Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics*, 4:1-13, 2015.
7. Christopher SG Khoo and Sathik Basha Johnkhan. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44(4):491-511, 2018.
8. Filipe Nunes Ribeiro, Matheus Arau´jo, Pollyanna Gon¸calves, Marcos Andr´e Gon¸calves, and Fabr´ıcio Benevenuto. Sentibench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5:1-29, 2015.
9. G. Palle R. Samuel, Rozzi. The dark side of sentiment analysis: an exploratory review using lexicons, dictionaries, and a statistical monkey and chimp. *Social Science Research Network*, 2022.
10. Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267-307, June 2011.