



Sentiment analysis and resources for informal Arabic text on social media

ITANI, Maher

Available from the Sheffield Hallam University Research Archive (SHURA) at:

<http://shura.shu.ac.uk/23402/>

A Sheffield Hallam University thesis

This thesis is protected by copyright which belongs to the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Please visit <http://shura.shu.ac.uk/23402/> and <http://shura.shu.ac.uk/information.html> for further details about copyright and re-use permissions.

Sentiment Analysis and Resources for Informal Arabic Text on Social Media

Maher Itani

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

August 2018

Abstract

Online content posted by Arab users on social networks does not generally abide by the grammatical and spelling rules. These posts, or comments, are valuable because they contain users' opinions towards different objects such as products, policies, institutions, and people. These opinions constitute important material for commercial and governmental institutions. Commercial institutions can use these opinions to steer marketing campaigns, optimize their products and know the weaknesses and/ or strengths of their products. Governmental institutions can benefit from the social networks posts to detect public opinion before or after legislating a new policy or law and to learn about the main issues that concern citizens. However, the huge size of online data and its noisy nature can hinder manual extraction and classification of opinions present in online comments. Given the irregularity of dialectal Arabic (or informal Arabic), tools developed for formally correct Arabic are of limited use. This is specifically the case when employed in sentiment analysis (SA) where the target of the analysis is social media content. This research implemented a system that addresses this challenge. This work can be roughly divided into three blocks: building a corpus for SA and manually tagging it to check the performance of the constructed lexicon-based (LB) classifier; building a sentiment lexicon that consists of three different sets of patterns (negative, positive, and spam); and finally implementing a classifier that employs the lexicon to classify Facebook comments. In addition to providing resources for dialectal Arabic SA and classifying Facebook comments, this work categorises reasons behind incorrect classification, provides preliminary solutions for some of them with focus on negation, and uses regular expressions to detect the presence of lexemes. This work also illustrates how the constructed classifier works along with its different levels of reporting. Moreover, it compares the performance of the LB classifier against Naïve Bayes classifier and addresses how NLP tools such as POS tagging and Named Entity Recognition can be employed in SA. In addition, the work studies the performance of the implemented LB classifier and the developed sentiment lexicon when used to classify other corpora used in the literature, and the performance of lexicons used in the literature to classify the corpora constructed in this research. With minor changes, the classifier can be used in domain classification of documents (sports, science, news, etc.). The work ends with a discussion of research questions arising from the research reported.

Keywords: opinion mining, sentiment analysis, social media, Facebook, Arabic language.

Acknowledgement

I would like to express my deep gratitude to my advisors Dr. Chris Roast and Dr. Samir Al-Khayatt; I could not have done it without their support. I want to thank them for their valuable advices since day one.

I would also like to thank my family: my wife, my mother, and my siblings, for their continuous support and motivation.

A special thanks to my friend Mohammad Sioufi for his technical advices and support.

Finally, I would like to thank the SHU administrators and specifically Ms. Rachel Finch and Dr. Francis Slack for their prompt support and help in admission and registration logistics.

Table of Content

Abstract.....	2
Acknowledgement	3
List of Tables	8
List of Figures	8
CHAPTER 1: Introduction	11
1.1 Background and Rationale	11
1.2 Research Motivation	13
1.3 Research Objectives.....	14
1.4 Research Questions	14
1.5 Research Methods	15
1.6 Contributions	15
1.7 Dissertation Outline	16
Summary	17
CHAPTER 2: Context and Literature	18
2.1 Introduction	18
2.2 Research Context	18
2.2.1 Data Mining.....	18
2.2.2 Association Rules	19
2.2.3 Supervised Learning.....	19
2.2.4 Evaluating Classifiers.....	20
2.2.5 Decision Trees	22
2.2.6 NB Classifiers	22
2.2.7 Support Vector Machines	23
2.2.8 k-Nearest Neighbours (kNN)	24
2.1.9 Unsupervised Learning.....	24
2.2.10 Web Mining.....	24
2.2.11 Information Retrieval.....	25
2.2.12 Evaluation	26
2.2.13 Pre-processing.....	26

2.2.14 Crawling	26
2.2.15 Social Media	27
2.2.16 Content, Usage, and Structure Mining	27
2.2.17 Sentiment Mining.....	28
2.2.18 Sentiment Classification.....	28
2.2.19 Feature-Based Sentiment Classification	28
2.2.20 Comparative Opinion Mining.....	29
2.3 Sentiment Analysis Literature	29
2.3.1 Sentiment Analysis of Social Media	30
2.3.2 Arabic NLP	33
2.3.3 Arabic Sentiment Mining Literature	34
2.4 Negation Literature	43
Summary	46
CHAPTER 3: Research Philosophy	48
3.1 Introduction	48
3.2 Types of Research Approaches	49
3.2.1 Deductive vs. Inductive Approach	49
3.2.2 Qualitative vs. Quantitative	50
3.3 Data Collection Technique	50
3.4 Data Analysis.....	51
3.4.1 Research Approach	52
3.5 Ethical Considerations.....	52
3.5.1 Intellectual Property	53
3.5.2 Copyright.....	53
3.5.3 Copyrights and social media	54
3.5.4 Facebook's Privacy settings	56
3.5.5 Data Management	56
3.5.6 Data Collection and Ethical Issues	57
3.5.7 Ethics Scrutiny	57
Summary	58
CHAPTER 4: Development of Sentiment Resources and Sentiment Classifier	59
4.1 Introduction to Classification using Lexicon	59
4.2 Building the Corpus.....	60

4.2.1 Current Corpora	60
4.2.2 Data Collection	68
4.2.3 Pre-processing	68
4.2.4 Manual Classification	69
4.2.5 Corpora Characteristics	70
4.3 Sentiment Lexicon	71
4.3.1 Importance of Sentiment Lexicon	71
4.3.2 Building the lexicon	71
4.4 Negation	78
Summary	80
CHAPTER 5: Classifier Design	81
5.1 Introduction	81
5.2 Classification Algorithm	81
5.3 Uploading Corpus, Lexicon, and Inverters	84
5.4 Classification Results and Statistics	84
Summary	87
CHAPTER 6: Analysis and Validation	88
6.1 Primary Results	89
6.2 Analysing the Results of Same-Domain Setups (NC-NL and AC-AL)	91
6.2.1 Neutral Errors	92
6.2.2 Negative Errors	96
6.2.3 Positive Errors	100
6.3 Analysing the Results of Cross-Domain Setups (NC-AL and AC-NL)	100
6.4 Effect of Increasing the Lexicon Size	102
6.5 NB Classifier versus LB Classifier	103
6.5.1 Same-Domain LB Setups versus NB Cross Validation Setups	104
6.5.2 Cross-Domain LB setups versus NB Train/Test Setups	104
6.6 Classification Results of Different Lexicons	105
6.7 Classification Results of Different Corpora	106
6.8 Spam Analysis	107
6.9 Negation Analysis	108
6.10 Domain Comparison	111
Summary	112

CHAPTER 7: Conclusion and Future work	113
7.1 Conclusions	113
7.1.1 Investigating Arabic Sentiment Analysis	114
7.1.2 Constructing the Corpus, the Lexicon, and the Classifier	114
7.1.3 Comparing LB and NB classifiers	115
7.1.4 Reasons of Incorrect Classification	115
7.2 Contributions	116
7.3 Future Work	116
References	118
Appendix A: Samples of Data and its Translation	140
Appendix B: Ethics.....	141
Appendix C: Research Data Management Policy	151
Appendix D: Data Management Plan.....	154
Appendix E: Classifier Design	156

List of Tables

Table 1 - Classification Confusion matrix	21
Table 2 – Level, Methods, Domains, Data Sources, and Features Used in Sentiment Analysis.	30
Table 3 – Examples of Different Corpora and their Properties	62
Table 4 - Source of Corpora mentioned in table 3	64
Table 5 - Frequency of Comments of Each Class.....	71
Table 6 - Numbers of Lexemes in the Lexicon Grouped Per Source	76
Table 7 - Percentage of each Dialect in the Lexicon	78
Table 8 - Percentage of Lexemes in Dialects.....	78
Table 9 – Common MSA Inverters.....	79
Table 10 - Common IA Inverters.....	80
Table 11 - Classification Truth Table	84
Table 12 - NB Classification Results.....	90
Table 13 - LB Classification Results of Initial Setups.....	90
Table 14 - Percentage of Errors of Each Category	92
Table 15 - Different Reasons Leading to Neutral Error in AC-AS and NC-NS	92
Table 16 - Different Reasons Leading to Negative Errors in AC-AS and NC-NS	97
Table 17 - Percentage of Different Errors in AC-NS and NC-AS	101
Table 18 - Effect of Increasing the Number of Lexemes on Performance of the Classifier	103
Table 19 - LB Classification Results	104
Table 20 - Results of Classifying AC and NC using Different Lexicons	105

List of Figures

Figure 4.1 - Sample of Downloaded Comments.....	69
Figure 4.2 - Sample Output from MADAMIRA's NER	73
Figure 4.3 – POS Tagging Sample Output 1	73
Figure 4.4 – POS Tagging Sample Output 2	73
Figure 6.1 - Sample Output from MADAMIRA's NER	92
Figure E 1 - Form Used to Upload Lexicon, Inverters, and Comments to be Classified.....	156
Figure E.2 - Form Used to Load Corpus, Inverters, and Lexicon.....	156
Figure E.3 - Form Used to Specify Number of Records.....	156
Figure E.4 - Sample of Classification Results	157
Figure E.5- Sample of Grouping Results	157
Figure E.6 - Sample of Filtering Results.....	157
Figure E.7- Sample of Classification Summary.....	157
Figure E.8- Frequencies and Percentages of Comments of each Class	158
Figure E.9 - Summary Showing Frequency of Lexemes	158

List of Abbreviations

AC: Arts corpus consists of 1000 Facebook comments. Its development is explained in chapter 4.

AI: Artificial intelligence

AL: Arts lexicon consists of three sets of sentimental lexemes extracted from AC.

AL2: represents the number of lexemes extracted from Anew.

Anew: represents a corpus of arts comments used to extract sentimental lexemes. It consists of comments other than those found in AC.

API: Application Programming Interface

ANLP: Arabic Natural Language Processing

ABSA: Aspect Based Sentiment Analysis:

CNew: A corpus consisting of 1000 arts and news comments and different from AC and NC. It is used to test the performance of the classifier on unseen corpus.

DA: Dialectal Arabic is the Informal Arabic used in daily verbal communication and is not governed by spelling and grammatical rules like Classical Arabic. Dialects differ depending on geographical location.

DT: Decision tree

FB: Facebook

FTP: File Transfer Protocol

HTML: Hyper Text Mark-up Language

IA: Informal Arabic refers to any dialectal version of Arabic without focus on a specific dialect.

IAA: Inter Annotator Agreement

IP: Intellectual Property

IR: information retrieval

KDD: Knowledge discovery in databases

kNN: k-Nearest Neighbours

ML: machine learning

MSA: Modern Standard Arabic

NB: Naïve Bayes

NC: News corpus consists of 1000 Facebook comments. Its development is explained in chapter 4.

NER: Named Entity Recognition

NL: News lexicon consists of three sets of sentimental lexemes extracted from NC.

NL2: represents the number of lexemes extracted from Nnew.

NLP: Natural Language Processing

Nnew: represents a corpus of news comments used to extract sentimental lexemes. It consists of comments other than those found in NC.

PMI: Pointwise Mutual Information

POS: Part of Speech

RE: Regular expression

SA: Sentiment Analysis

SM: Social Media

SNT: NS plus additional lexemes extracted from news posts

SSA: Subjectivity and sentiment analysis

SVM: Support Vector Machines

URL: Uniform Resource Locator

WIPO: World Intellectual Property Organization
WWW: World Wide Web

CHAPTER 1: Introduction

1.1 Background and Rationale

“What Women Want” is a 2000 American romantic comedy movie written by Josh Goldsmith that describes adventures encountered by a marketing executive who accidentally gets the power of reading women’s mind; using this “superpower”, he becomes able to craft the best marketing strategies for his company’s products.

Despite the romantic and comic course of the movie, it addresses an important issue, which is the significance of knowing what people feel. Decisions are affected by opinions: knowing what others feel towards an object (product, policy, organization, candidate, etc.) can affect decision-making. We tend to believe what the majority feel or say towards something. If many people recommended a restaurant for us, for example, we will most likely have a positive feeling towards the restaurant. This applies to many other topics. The fast growth of the World Wide Web (WWW) provided the medium needed to express opinions and to know the opinions of others. Web 2.0, a term first coined by Dinucci (1999), was formally introduced by O’Reilly (2005) with user-driven content being the most significant feature. It marked the era where most websites have rating features that allow a customer or a client to express an opinion about an object or a service. Some sites also allow users to post textual data that express their opinions. These opinions are important for many reasons: the owners of the website can know what others think of their products (or any other object like a candidate or a policy) and changes may be made accordingly. For example, if many users of a certain mobile phone complained about the battery life, this is considered an indicator that a problem exists and an action to fix it should be taken. On the other hand, if a potential customer is searching for a new laptop, he or she may be influenced by other people’s feedback about a certain brand or model and buy it. If an organization (political or commercial) knows what people feel towards something, marketing can be made to target potential customers’ needs, and specific advertisements can be created to guarantee catching the customers’ attention.

Similarly, if an organization knows what others are complaining about, it can launch new products and policies to satisfy the targeted audience. For example, if a mobile phone company

X knows that customers of another mobile phone company Y are complaining of low camera resolution, then company X can launch a marketing campaign with a focus on the camera resolution. If a governmental organization is willing to enact a new policy, knowing what people think about the policy before it is applied may help in applying the policy in the proper settings with modifications based on people's opinions. For all these reasons and many others, it is vital to know people's opinions. However, the scale of the task of assessing opinions is of great significance - the number of online users has increased tremendously to reach 2.2 billion in 2016 and is expected to rise to 2.72 billion in 2019 ("Number of Social Media Users", 2016). This makes manual extraction and classification of opinions an infeasible task, and an automated process is needed that can classify comments present in a large dataset (corpus).

In addition, the WWW is currently involved in all aspects of life: education, advertisement, business and other fields depend on the WWW because of its availability, simplicity, and ability to facilitate plenty of services with simple clicks. Moreover, communication and sharing of ideas are now easier because of the user-friendly interfaces that the WWW provides. The improvement in network technologies, and specifically the Internet, has allowed users to share different types of media (text, audio, and video) in a simple and mostly free manner. The implementation of WWW adopts the client-server model, where users, using client programs (such as Telnet, SSH, or FTP client) can have access to data hosted on servers. Roughly speaking, the WWW has the following characteristics ("World-Wide Web", 2016):

- a) The size of online data is huge and continuously growing.
- b) The online data are of different types: images, text, audio and video.
- c) Backbone of social networks: The WWW hosts different online societies of different domains such as chat rooms.
- d) Web services: Commercial, educational, governmental and other services are now available through the web.
- e) Online data have a noisy nature: Almost all pages, regardless of their content, have noisy data such as banners, headers and footers, and advertisements that may not be related to the main content of the page.

- f) Redundancy and discrepancies: Since there is large number of authors, same content may be hosted online by different authors. Moreover, false data and incorrect content are also present in huge amounts.
- g) Dynamicity: Due to change in policies, customers' tastes and other elements, online data are prone to continuous changes in style and content.

1.2 Research Motivation

We chose to work on Arabic language for the following reasons:

1-The significant number of online users which was estimated to be 2.2 billion in 2016 and expected to rise to 2.72 billion in 2019 ("Number of Social Media Users", 2016), and therefore manual extraction and classification of comments written by these users cannot be done manually and need to be automated. Moreover, the huge number of Arab users indicates that there are many potential institutions (both governmental and commercial) that would benefit from the presence of a system that can extract and classify online comments according to their sentiment.

2-Although Arabic Natural Language Processing (NLP) has improved significantly in the last two decades, it is considered under-resourced when compared to English language, and thus the research community may benefit from additional resources such as annotated corpora and sentiment lexicon.

3-This work covers social media, and more precisely, textual data written in Dialectal Arabic (DA) and posted on Facebook. DA will be referred to as Informal Arabic (IA) hereafter. The number of Facebook users worldwide exceeded 2.2 billion in April 2018 (Most famous social network sites 2018, 2018), approximately 141 millions of whom are Arabs (Arabic Speaking Internet Users Statistics, 2017), who speak different dialects. The third motivation behind this work is to study the effect of the irregularity of IA on SA; irregularity of IA includes spelling, grammar, and style of writing.

In this work we refer to DA by IA because we do not differentiate between dialects neither do we study the association between a specific dialect and the sentiment of a comment. Moreover, none of the developed resources is dialect-specific.

1.3 Research Objectives

Due to all the characteristics mentioned earlier (mainly the dynamism and complexity of the Internet textual data) a potential exists for a system that can extract and classify the opinions present in these data. This work approaches this challenge and provides annotated resources (corpora and lexicon) to be used by the research community.

The two main obstacles that face effective classification when dealing with IA are (1) the limited number and accuracy of tools such as morphological analysers, Part of Speech (POS) taggers, stemmer, etc., and (2) the scarcity of tagged corpora that can be used to conduct experiments and the limited research done in this area when compared to what has been done for the English language. There are different types of classifiers, and these types will be briefly described in section 2.1. However, regardless of which language they address, they rely on the grammatical and spelling rules of the language. This characteristic makes them of no use when dealing with dialects that do not follow such rules. On the other hand, one specific type of classifiers, LB classifiers, classifies a sentence or a document depending on the semantic polarity or orientation of its words and phrases. Such classifiers are flexible because they allow for easy maintenance and allow updates to be made so the classification system can be applied in different domains (politics, sports, news, etc.). The proposed research objectives are to:

- Investigate (identify) classical techniques used in sentiment analysis (SA) with focus on Arabic language.
- Implement an LB sentiment classifier to classify social media (SM) comments written in IA and investigate how it can provide a better understanding of SA of IA.
 - Construct an annotated corpus (large collection of text) to be used for SA.
 - Construct an opinionated lexicon (a dictionary that assigns a polarity (positive, negative, etc.) to words instead of meaning)
- Compare the performance of an LB classifier with other Machine Learning classifier such as Naïve Bayes (NB) classifier.
- Identify main reasons behind incorrect sentiment classifications

1.4 Research Questions

Following the rationale mentioned above and the research objectives, the approach developed in this work provides a potential solution that is not dialect-specific and thus can be applied to IA.

Three main questions arise:

1. How can we get better understanding of SA of SM comments written in IA?
2. How can we improve sentiment classification of SM comments?
3. What are the main reasons behind incorrect classification of IA when an LB classifier is used?

1.5 Research Methods

Research methods can be roughly categorized into two main categories: quantitative methods that deal with well-defined metrics for success and failure and qualitative methods that deal with poorly structured data and try to interpret what they mean. They can also be categorized as deductive methods that are top-down approaches, which start with solid theory and try to narrow the research down to come with hypotheses to test the theory, and as inductive methods that are bottom-up approaches, which start from observations and poorly structured data and try to formulate a specific pattern or behaviour. To answer the research questions, we follow quantitative/deductive methodology: roughly speaking, we try in this work to classify Facebook textual comments of two domains (arts and news) as positive, negative, spam, dual and neutral. A manually built semantic lexicon is used, which contains opinionated words to be used in classification. Afterwards, we address the different categories of reasons that led to incorrect classification of comments such as misleading patterns, sarcasm, and negation. This work provides additional annotated resources to be used in SA of IA. It describes how the resources were constructed and used, it uses different ML tools along with the LB classifier, and it analyses different reasons behind incorrect classification and provides potential solutions to some of them.

1.6 Contributions

Itani et al. (2012) explain a comparison between an LB classifier and an NB classifier. Their initial results show that the LB classifier outperformed the NB classifier. Itani (2017) provides an annotated corpus of Informal Arabic texts available for public use. The corpus contains 2000 FB comments written in Informal Arabic and annotated using five labels: positive, negative, dual, neutral, and spam. Itani et al. (2017a, 2017b) explain the procedure of developing the sentiment resources for Informal Arabic. Specifically, these works describe how their corpus and lexicon were constructed and annotated. Chapter 5 provides one approach on how an LB classifier can be designed. Chapter 6 provides detailed analysis on the categories of errors encountered during

classification and discusses possible resolutions to the errors. This work also compares classification results of LB and NB classifiers. It studies as well how different NLP tools were used in SA context. Finally, this research study evaluates the developed corpora and lexicon by conducting several setups in which the constructed lexicon is used to classify corpora that are used in the literature. Other setups use different lexicons used in the literature to classify the developed corpora.

1.7 Dissertation Outline

Chapter 2 consists of two parts: context background and literature review. The context background aims to introduce the main terminologies and concepts that the reader needs to know before reading the literature review and other chapters. It starts by introducing the platform of data mining, and then it introduces different schools, algorithms, and definitions. The literature review discusses similar works and different techniques adopted and ends by summarizing the limitations that hinder sentiment classification, specifically for Arabic language.

Chapter 3 discusses the research methodology followed and how it adheres to the aims and objectives of the present work.

Chapter 4 describes, in chronological order, the phases followed in our work and the building blocks of our classifier. It explains how the corpus and the lexicon were built, and how a specific feature of a Regular Expression (RE) was used in the construction process.

Chapter 5 shows implementation details and how different pieces are put together. It also describes the user interface of the implemented system.

Chapter 6 discusses the validity of the proposed approach when compared against the literature. It provides detailed analysis of results and compares them to results ML classifier, including categories of errors encountered, suggests an approach to resolve negation, and suggests a potential resolution to sarcasm.

Chapter 7 concludes our outcomes along with the remaining limitations, provides recommendations, and identifies future goals. The detailed dissertation outline can be found in the table of content.

Summary

Chapter 1 explains how the growth of WWW and social media specifically provided a platform for online users to express their opinions towards different objects. It explains why these opinions are important for decision makers and the infeasibility of manual classification of the sentiment of these opinions. The chapter specifies the research motivations, objectives, and the questions. It also specifies the main contributions of this work: (1) creating new resources for SA of informal Arabic, (2) using an LB classifier to classify textual data, (3) analysing the reasons behind incorrect classification, (4) comparing LB to NB classifiers, (5) comparing the performance of developed lexicon when used to classify different corpora, and (6) comparing the results of using lexicons used in literature to classify developed corpora. The chapter ends by outlining the remaining of the dissertation.

CHAPTER 2: Context and Literature

2.1 Introduction

This work is one approach towards building an LB classifier that is not dialect-specific, in addition to providing new resources (corpora and lexicon) for SA of IA. Given its major effect on SA, we also studied the effect of negation on SA. Additionally, we studied the performance of our lexicon when tried on different corpora, as well as trying other lexicons on our corpora. We also compared our LB classification results against NB classification results.

Sections 2.2 briefly explains some paradigms that may be used in the research mentioned in the literature review, and not because they were adopted in this work. The topics include data mining basics and illustrate the use of association rules. They also cover the means by which a classifier is evaluated, the most commonly used supervised learning classifiers, and a summary about unsupervised learning. Afterwards, section 2.3 provides literature review related to SA in general, SA in SM, Arabic, NLP, Arabic SA, and negation. Section 2.4 covers negation literature.

2.2 Research Context

2.2.1 Data Mining

Knowledge discovery in databases (KDD), also known as data mining, is used to identify useful information hidden in data. Data mining can be applied to texts, images, databases, online webpages and other sources. Data mining employs NLP, ML, artificial intelligence (AI), mathematics, information retrieval (IR) and other fields. The major mining processes are association rule mining (to discover interrelation between variables within a data source), supervised learning (also known as classification), and unsupervised learning (also known as clustering). Data mining usually starts by pre-processing in which noisy parts of data are taken out. For example, if data mining is being applied on web pages, data mining starts by taking out unnecessary data such as Hyper Text Mark-up Language (HTML) tags, timestamps and other irrelevant data. Afterwards, a data mining approach is used to operate on raw data and produce useful knowledge. The last step in data mining is to evaluate the quality of knowledge extracted to see whether they are useful or not. Classical data mining uses structured data such as those stored in relational databases. On the other hand, and due to the quick growth of WWW, a new

branch of data mining, web mining, attracted many researchers due to the importance of data available online.

2.2.2 Association Rules

The aim of association rules is to find correlations connecting data components of a data source, such as tables of a database, or fields of a table (Agrawal et al., 1993). One typical example is extracting association rules governing the items bought from a supermarket. For example, we may find out that in 90% of the times, when chips are bought, a soft drink is bought. Such information can be used in placing the two products next to each other to increase the sales of both. Briefly, the concept of association rules can be summarized as follows: given a set I of items and a set T of transactions where each element in T is a set of items subset or equal to I , an association rule can be represented as follows:

$A \rightarrow B$, given that A is a subset of I , B is subset of I and $A \cap B = \emptyset$

A and B are called item sets. In other words, association rules' aim is to find all rules in set of transactions T that have specific values of support and confidence, where support is the ratio of transactions containing $A \cup B$, and confidence is the ratio of transactions in T containing A that contains B (Agrawal et al., 1993). Different algorithms exist for finding these rules such as Apriori algorithm, PrefixSpan algorithms and others.

2.2.3 Supervised Learning

The intuition behind this kind of mining is to learn new knowledge based on previous experience, where an experience is represented as computer data records (Caruana and Niculescu-Mizil, 2006). Each record is described by a set of features or attributes and one of these attributes is considered a target attribute or class. The aim of a supervised learning process is to create a classifier that can find a relation between attributes and the class in order to be able to predict the class when given unseen records where the class value is unknown. In other words, the classifier will learn from a set of examples a function that relates the attributes to the class.

The data used by the classifier to learn the relation between attributes and the class is known as training data ($Data_{train}$). Afterwards, when a relation is found, the classifier is fed with another set of unseen data, also known as test data ($Data_{test}$), to check the efficiency of the learning process. For this process to be successful, the test data should not be used in the learning process, and in order to check the efficiency of the classifier, the class of the test records should be known so

that they can be compared against the class predicted by the classifier. One example would be weather prediction. Given the temperature, humidity, wind speed, and the class “will rain”, we need to train a classifier that will learn from some records and finds a relation between values of fields and the value of the class. Afterwards, a different set of records will be used to check whether the classifier is able to predict the class of records. The efficiency of classification can be measured by computing the percentage of correct predictions out of total number of test records.

2.2.4 Evaluating Classifiers

When a classifier is created, its accuracy should be tested before deploying it. We mentioned earlier that test data could be used for this purpose, by dividing the number of correctly classified instances (of the test data) by the total number of instances. This measure is known as accuracy. When comparing performance of different classifiers, we usually compare their accuracy when given the same classification task, i.e., when given the same training and test data.

Given processed data (data ready to be input to a classifier), the data is split into two parts: $Data_{train}$ and $Data_{test}$. The size of each set depends on the overall all size of data and the way in which data is collected. If data collection is an on-going process, the data collected earlier can serve as training data, and the ones collected later will serve as test data (under the assumption that there is no significant change over time.). Cross validation offers a useful approach to increase confidence in learning results. The data set is divided into n distinct sets, $n-1$ of these sets will be used for training and the remaining set will be used for testing. The process is repeated n times by changing the $n-1$ sets used for training and the set used for testing, average accuracy is then used to evaluate the classifier. As mentioned earlier, test data should not be seen by the classifier during training. 10-fold cross validation is commonly used.

In some classification tasks, we need to know whether a data record has a specific class or not. In such binary classification, the class which we are interested to detect is called positive class; the other classes are called negative classes. Usually in such cases, our class of interest is a minority among total instances. For example, if we are classifying online email registration requests as legitimate or not, and assuming that the majority of these requests are legitimate, using the accuracy as a measure would be misleading since it does not reflect the efficiency of the classifier. Assume that in 3 out of 100 instances, the request is fake, so by classifying all requests

as legitimate, the accuracy will be 97% without actually classifying anything or detecting the class that it was supposed to detect. That is why more effective evaluation criteria are considered such as the F-measure that is used mainly in SA context (Agarwal et al., 2011; Abdul-Mageed and Diab, 2014, Korayem et al., 2012).

The F-measure is more precise in evaluating classifiers. It depends on two parameters: precision and recall. Both parameters are used in a confusion matrix that shows results predicted by classifiers and actual results (test data have known class values to be used in evaluating the classifier). The confusion matrix consists of four entries: true positive (TP), true negative (TN), false positive (FP), and false negative (FN):

TP: number of correctly classified positive instances

TN: number of correctly classified negative instances

FP: number of incorrectly classified negative instances

FN: number of incorrectly classified positive instances

Table 1 shows the confusion matrix that relates the four parameters mentioned above (TP, TN, FP, and FN):

Table 1 - Classification Confusion matrix

	Actual Positive	Actual Negative
Classified as Positive	TP	FN
Classified as Negative	FP	TN

After defining the four classification possibilities (TP, TN, FP, FN) and computing their values, they can be used to determine the values of precision (P), recall (R), and F1-measure according to the following formulas:

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

Equation 1-Precision

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

Equation 2-Recall

$$F1 = \frac{2 * P * R}{P + R}$$

Equation 3-F1-Measure

It is worth mentioning that other variants of this formula exist such as F2-measure that gives higher weight to recall than precision or F0.5-measure that gives higher weight to precision than recall. However, in data mining context, F1-measure is the most commonly used formula (Doreswamy, 2012). When dealing with binary classification, an average F-measure can be used when more than two classes are available, each time setting the target class as the positive class, and all the rest combined as the negative class. The average F-measure is the measure adopted in this work.

2.2.5 Decision Trees

One other supervised learning approach is decision trees. Its high accuracy and ease of implementation makes it one of the most commonly used classifiers. Each node in the tree represents a test (like an if-statement) of one feature of the data record, leaves of the tree represents the class of each branch of the tree given the values of the tests at each node. The main algorithm used to build decision trees is called ID3 and was introduced by Quinlan (1987). ID3 typically uses greedy search algorithm.

2.2.6 NB Classifiers

NB classifier is a probabilistic classifier that assumes independence of attributes (Doreswamy, 2012). Given a data set D , let the attributes x_1 through x_n represent attributes of each record in D . Let C represent the set of values c_1 through c_k of the class attribute. Given an instance y with a_1 through a_n as values of attributes, the NB classification will select c_i with highest probability according to the following formula:

$$\Pr(C = ci | x_1 = a_1, \dots, x_n = a_n) = \frac{\Pr(x_1 = a_1, \dots, x_n = a_n | C = ci) \Pr(C = ci)}{\Pr(x_1 = a_1, \dots, x_n = a_n)}$$

Equation 4-NB Posterior Probability

Assuming the independence of attributes is invalid in most applications, however, results achieved by many researchers show that the NB classifier is efficient in text classification despite the invalidity of the assumption of independence of attributes (Rish, 2001).

2.2.7 Support Vector Machines

Support Vector Machines (SVM) classifiers are among the most efficient classifiers when the number of attributes is high (Amancio et al., 2014). SVM is a linear learning approach that uses binary classifiers. The main intuition behind SVM is to set a boundary between positive and negative instances. To do this, it finds a function $g(x)$ (no need to know what the function is when using SVM) that classifies x (where x is the input vector with n attributes) as positive if $g(x)$ is non-negative and negative otherwise.

Graphically, the hyperplane created by the function will split the input into two parts, one containing the positive instances and the other containing the negative instances. The line that corresponds to the linear function found by SVM separates negative and positive instances.

In short, SVM, aims to find a maximal margin decision boundary that separates the two classes. If the two classes cannot be linearly separated, the boundary is found by transforming the input space (all instances of the data set D) into an n -dimensional space instead of a hyperplane and the separation becomes a plane instead of a straight line. Nonetheless, there are a few drawbacks to be considered when SVM is used:

- SVM operates in real space, so if the attributes are not numeric, they should be converted to numbers before SVM can be used. This can be done by representing each attribute by another attribute of Boolean value that will be 1 if the attribute exists and 0 if it does not.
- SVM can be used as a binary classifier. If more than two classes exist, SVM cannot be used directly, and major modifications should be applied before using SVM.
- The hyperplanes created by SVM are nontrivial and understanding them visually is a hard task for humans.

2.2.8 k-Nearest Neighbours (kNN)

Decision trees, NB, and SVM classifiers learn models from training data then apply this knowledge on test data. kNN is different in that it does not learn models from training data. It only learns a model when trying to classify a test data. Given a training data D , kNN will not use records of D to learn models. It will compare each instance d of the test data with instances present in D , and then it will check the similarity between d and every record of D . kNN will then assign to d the most frequent class that occurred in neighbours of d . The core component that will determine the efficiency of this approach is the selection of the function to be used when computing similarity. The function may be classical Euclidean distance. Some researches (Yang and Liu, 1999) claimed that the kNN can be as efficient as SVM classifiers.

2.1.9 Unsupervised Learning

Unlike supervised learning that tries to find a relation between values of attributes and the class attribute, the class attribute does not exist in some applications. In other words, in supervised learning there are input variables (a) and an output variable (b), and the objective is to map the input to the output. On the other hand, unsupervised learning aims to model the distribution in the data in order to learn more about it. In such cases, the unsupervised classifiers tend to divide the input space into clusters based on similarity among instances with each cluster including similar instances. Each record can be thought of as a point in n -dimensional space where n is number of attributes of each data instance. Similar to the kNN classifier, a function is needed to check whether instances are similar to each other (this similarity is the distance separating two points, where each point is a data instance). The choice of the similarity functions depends on the nature of data being clustered, specifically whether the attributes are numeric or nominal. One typical application where clustering can be used is to categorize a set of documents according to their similarity: sports, arts, news, etc.

2.2.10 Web Mining

The main difference between data mining and web mining is the nature and structure of data. This difference leads to a change of algorithms and tools used in both cases. IR aims to retrieve documents that fit a query submitted by a user. For example, if the user query is “how to make hot chocolate”, the objective is to retrieve documents that are relevant to the keywords of the

query, and thus help answer the query. Efficiency of IR systems is assessed by the precision and recall of retrieved documents given the user's query statement.

2.2.11 Information Retrieval

In IR context, the document is considered to be the smallest unit of data. The major objective of IR is to retrieve set of documents from a bigger set given a query. For instance, consider a set of 1000 documents covering 3 topics: arts, sports, and news. If the user wants to retrieve only documents related to sports then a query containing keywords (such as new, space, and technology) may be used for this purpose. The effectiveness of the IR system is then measure by the relevance of retrieved documents. Consider a search engine; if the user entered a search phrase such as "best coffee shops in France", millions of documents may be retrieved, however, the order in which the relevant pages are ranked is what make a search engine better than others (Frakes and Baeza-Yates, 1999; Baeza-Yates and Ribeiro-Neto, 1999; Grossman and Frieder, 2012; Büttcher et al., 2016).

A model in IR context defines how queries and documents/pages are to be represented and specifies the criteria to determine which documents are relevant to the query. The three main models are language model, Boolean model, and vector space model. These three models consider the documents and queries as "terms" or "bag of words" regardless to their sequence in the documents and queries.

Language Model: This model depends mainly on probability. For each candidate document, the documents are ranked according to the likelihood of the query being relevant.

Boolean Model: each term in the document will have a weight of 0 or 1 if it contains terms from the query. Each document is then represented by a vector where each term, or word, in the document is either present or not. Documents are then retrieved if there is an exact match between the query and the query. Logical operators (And, Or, Not) can be used to limit the number of retrieved documents.

Vector space model: In this model, each term in the document is given a weight, not necessarily 0 or 1. Several variations of his model exist; a document is retrieved according to their relevance to the query.

2.2.12 Evaluation

The concepts of precision and recall as used in ML (see 2.1.4) are also used in IR, since the retrieval can be viewed as selecting documents in a class matching the user's query. A precision-recall curve and confusion matrix can also be used (Davis and Goadrich, 2006).

2.2.13 Pre-processing

Common pre-processing techniques include removing stopwords and stemming. Stopwords are frequent words that do not contribute to the selection process, words such as articles, pronouns and prepositions. Stemming means normalizing different syntactical variants of a word to the main stem; this will decrease the number of words in the documents and improve the performance of the models. Pre-processing also includes removing redundancies, numbers, and other word types that are considered irrelevant.

For the specifics of web mining, html tags are considered irrelevant and are removed from the document prior to applying the model; some of these tags however may help to specify which parts of the text are more important than others.

2.2.14 Crawling

Given the huge amount of online data, programs are needed that can automatically download data, this can be done using crawlers (Pant et al., 2004). Crawlers are automated processes that are used to visit online pages, download them, and store them in some repository. Crawlers then use the links in visited pages to identify which other pages to visit. Since these pages are not static, crawlers are designed to cope with the dynamic nature of online content. One of the basic usages of crawlers is business intelligence where companies can collect data posted on competitors' websites, another usage would be to automatically collect email addresses online to use them later on for marketing purposes; crawlers can also be used to prepare corpora. Crawling starts by visiting a root uniform resource locator (url) then it visits hyperlinks present on the page and download contents of target pages. The process continues until all pages have been downloaded or until the target number of pages has been reached. Although implementation details of commercial crawlers cannot be known, different theoretical aspects of implementing crawling algorithms were addressed by different researchers (Pant et al., 2004; De Bra and Post, 1994; Chakrabarti et al., 1999; Cho et al., 1998; Cho and Garcia-Molina, 1999).

2.2.15 Social Media

Social media constitute friendly web-based platforms to socialize and share opinions. The number of users of social networks exceeded 2 billion users (“Social Networking Statistics”, 2014) with 1.4 billion using Facebook. The advancement in mobile phones technologies and tablets contributed to the rapid growth of these networks. Social networks can be roughly divided into seven categories (White, 2014): academic (such as Academia.edu), professional (such as LinkedIn), multimedia sharing (such as Youtube and Flickr), social connections (such as Facebook and Twitter), educational (such as The Student Room), informational (such as Do It Yourself Community), and hobbies (such as Oh My Bloom). The growth of social networks and its social and political effects were studied in (Backstrom et al., 2006; Haythornthwaite, 2005; Trusov et al., 2009).

2.2.16 Content, Usage, and Structure Mining

Web mining extracts useful information from the unstructured data of webpages. This knowledge may be part of the webpage content, structure or logs of usage. We briefly explain each of these aspects of web mining:

Content Mining: This is the closest to the classical data mining; it includes classifying different webpages according to their content (politics, sports, etc.) It also includes mining content of pages and classifies them according the sentiment present in them (negative, positive, etc.), or extracts any other target type of information from these pages. Due to the unstructured nature of these data, classical data mining techniques and database design approaches had to be modified to fit the new type of data (Mobasher et al., 2000; Liu and Chen-Chuan-Chang, 2004; Shyu et al., 2007).

Usage Mining: this branch tends to discover patterns in which users browse a website: what do they focus on, which locations can be used to place ads, which sequence of clicks is followed by users, and in which sequence users may go from one webpage to another. This can be done by checking server logs to see the navigation sequence of users, such patterns of navigation can help in cross marketing; if we can know which page or parts of the pages attract users more than others, ads can be placed accordingly. Some users are interested in textual data whereas others are interested in multimedia. This can be known after analysing the usage patterns of users (Srivastava et al., 2000; Mobasher et al., 2000; Spiliopoulou, 2000).

Structure mining: given that webpages are connected via hyperlinks, this branch includes discovering new webpages; this is the core concept behind web crawling. Moreover, structure mining help knowing the hierarchy of websites (Chakrabarti et al., 1999; Han et al., 2000).

2.2.17 Sentiment Mining

Different data mining models mentioned in previous sections may be used in sentiment mining. Online data has valuable information, which is users' sentiment towards an object (policy, product, etc.). This field attracted researchers for the last two decades since this information can help in decision making. The size and nature of data enforces automation of this process. If the marketing officer at an institution knows what consumers like, marketing campaigns can be steered accordingly. For example, if we know that a Facebook user is a fan of mobile brand x, placing ads about this product on user's page would be a good idea, at the same time, if we know that he dislikes this brand, then placing an ad for a competitor brand is a better idea. This work specifically addresses SA of social media comments written in dialectal Arabic.

2.2.18 Sentiment Classification

Besides classifying a text as subjective or objective, the polarity of subjective data should be known before it can be properly employed. Sentiment polarity can negative, positive, and dual. Different approaches have been tried, either applying sentiment classification on sentence level or document level (Hu and Liu, 2004; Dave et al., 2003; Farra et al., 2010; Hamouda and El-Taher, 2013; Pang et al., 2002; Hatzivassiloglou and Wiebe, 2000; Kim and Hovy, 2004).

2.2.19 Feature-Based Sentiment Classification

Objects such as products or policies have features towards which users may have different opinions. For example, for a product x, a user may have positive feedback towards some of its features and negative ones towards others. Feature-based sentiment classification zooms into the product to know which features are being commented on by users (whether negatively or positively) and to know the polarity of sentiments (Rohrdantz et al., 2012; Eirinaki et al., 2012; Pang and Lee, 2004; Hu and Liu, 2004). Hu and Liu (2004) were among the first to address feature-based opinion summarization. Association rules were used based to extract frequent product features. Unsupervised learning was used by Popescu and Etzioni, (2007), a system, OPINE, was developed for this purpose that extract and classify opinions from customers' review. Zhang et al., (2010) used product-based keywords to extract product features. Khan et al.

(2010) exploited a grammatical phenomenon, which is the use of auxiliary verbs in opinionated sentences, for this purpose. According to their experiments, auxiliary verbs were used in more than 80% of opinionated sentences. Zhai et al. (2010) suggested an LB approach that considers the structure of a review to enhance performance.

2.2.20 Comparative Opinion Mining

Besides extracting opinions and determining sentiments, online data may contain a comparison between entities or features of two entities. This is a more detailed and harder task to achieve since some features of an object x may be better than those of object y , and at the same time the opposite may be true, i.e., some features of object y are better than those of x . This problem has two sides, first to extract which features are being compared (and to which entity they belong to) and the sentiment of comparison. For example, the resolution of camera x is better than camera y , but at the same time, the battery life of camera y is longer than that of camera x . Comparing sentiments within reviews was addressed by Jindal and Liu (2006) and by Pang and Lee (2004).

2.3 Sentiment Analysis Literature

Hatzivassiloglou and McKeown (1997) were the first to study the sentiment orientation of words. Their approach depended on adjectives and conjunctions used in English language. Their approach was tried on Wall Street Journal. A log-linear regression model was used to check the orientation of couples of adjectives. The accuracy reported was 82%. Their approach starts by extracting adjectives connected with conjunctions (and, but, etc.). Then a supervised learning algorithm is used to cluster adjectives based on their similarities.

Turney (2002) used unsupervised learning to classify a review as positive or negative. His approach starts by extracting adjectives and adverbs, and then the semantic orientation is computed using PMI-IR. PMI-IR is used to measure the similarity of pairs of words and the orientation of a review is estimated as the average of semantic orientation of its phrases.

Hu and Liu (2004) addressed product reviews submitted by customers. Their work provides a summary of positive and negative opinions expressed about product features (battery life, phone size, etc.). Their approach starts by collecting reviews, feature extraction and pruning are then applied, opinionated words are then extracted and given a polarity, and finally a sentiment summary is generated. Feature extraction starts by deciding which features will be used in the classification process (stylistic, semantic, etc.), and features pruning takes out insignificant

features that do not affect classification performance. Work related to sentiment mining from product review was addressed in many research studies (Popescu & Etzioni, 2007; Morinaga et al., 2002; Lee et al., 2008; Goldberg & Zhu, 2006; Dave et al., 2003).

Linguistic rules discussed by Hatzivassiloglou and McKeown (1997) were enhanced and used by Kanayama and Nasukawa (2006). Classifying a document as positive, negative, or neutral can be done either by classifying the document as a whole, or by classifying it using the sentiment of its sentences (Kim & Hovy, 2004; Wiebe & Riloff, 2005; Wilson et al., 2004). Sentiment classification approaches can be roughly divided into two main categories: corpus based and dictionary based (Abdulla et al. (2013)). In corpus-based approaches, sentiment is determined by considering co-occurrence of opinionated words (Dave et al., 2003; Hatzivassiloglou & Wiebe, 2000). Dictionary based approaches depend on synonyms and antonyms (Hu & Liu, 2004; Wiebe & Riloff, 2005; Leacock & Chodorow 1998). Bhuiyan et al. (2009) introduced a comprehensive study of mining opinions from customer feedback. The authors evaluated the different techniques followed and categorized them according to their strengths and weaknesses. Table 2 provides a rough useful breakdown of SA methods, levels of classification, domains, sources of data, data source, and features used in classification. Two important surveys summarize studies that fall into categories mentioned in table 2 were conducted by Bhuiyan et al. (2009) and Abbasi et al. (2008).

Table 2 – Level, Methods, Domains, Data Sources, and Features Used in Sentiment Analysis.

Levels of Classification	Features Used	Method	Domain	Data Source
Document level vs Sentence level	Stylistic	ML	News	Forums
Subjective vs Objective	Semantic	LB	Arts	Social Media
Negative, Neutral, Mixed, or Positive	Syntactic		Economics	Website

2.3.1 Sentiment Analysis of Social Media

The number of social media users increased rapidly in the last few years. Some of them have over billion users (“Social Networks Statistics”, 2013). People join social media for many reasons (“10 reasons people use social media’, 2013). One of the top ten reasons is to express opinions and know the opinions of others about different topics (politics, commercial products, sports, etc.). The friendly and easy- to- use online websites allow users to express their opinions and share them with the public. Approximately 300,000 textual comments are posted every minute (Flacy, 2011).

Johnson et al. (2012) classify political sentiments present in tweets (a tweet is the term used to describe a post on Twitter). They specifically classify opinions expressed about the ex-US president Obama. Three different approaches were designed for this purpose: rule-based, supervised, and semi-supervised. The approaches were evaluated by checking the accuracy of 2500 tweets that were manually labelled. The first approach is LB; it searches tweets for presence of positive or negative words and phrases. Tweets were classified as leaning towards being positive or negative according to the number of opinionated words present in the tweet; the tie was broken by choosing a polarity (positive or negative) randomly. In supervised learning, a classifier learns different features contributing to the polarity of a manually classified tweet. Then the classifier is given a new unclassified tweet. Maximum Entropy classifier was chosen by the authors, which is a classical probabilistic classifier used for text classification. Unigrams, bigrams and emoticons are used as features for this classifier. In the third approach, Twitter label propagation graph was used, which uses a weighted graph whose vertices represent users and their tweets connected by weighted edges. Accuracy reported ranged between 30% and 69%. Barhan and Shakhomirov (2012) used SVM classifier and n-grams were used as features. To measure performance of classifiers, precision, recall and F1-measure were used.

Precision reported by the authors ranged between 0.62 and 0.65, and recall ranged between 0.71 and 0.76. The authors also revisited definitions of opinions present in tweets. Pak and Paroubek (2010) applied linguistic analysis on a corpus of tweets and the phenomena observed were used to build a classifier that classifies tweets as negative, positive and neutral. Tweets were searched for the presence of negative and positive lexicons, and these tweets were used to train the classifier presented by the authors. Objective tweets were collected from newspapers' Twitter accounts. The authors assumed that the presence of a lexicon is enough to give the tweets its sentiment since the tweet size was limited to 140 characters (tweet size increased to 280 characters in 2017). The authors reported that different classifiers were tried and that NB classifier gave the best results, POS tags and n-grams were used as features. Results showed high accuracy (precision) and low decision (recall).

Saif et al. (2012) used semantic features to train a tweets classifier. F1-measure of 75.95% was reported although stop words were not removed. Their results were 6.47% higher than baseline approach when unigrams alone were used and 4.78% higher when POS feature were used with

unigrams. Hamouda and Akaichi (2013) studied Facebook “statuses updates” written in English and posted by Tunisian users during the Arabic Spring. The objective of this research is to analyse the social behaviour of Tunisians during a critical event and whether textual data can be used to know the public opinion during that event. Status updates were collected from randomly selected Facebook users of different ages, genders, occupations and social statuses, and two ML classifiers were used, namely SVM and NB. Their approach consists of 5 phases: collection of comments, creating sentiment lexicons, pre-processing, feature extraction, and classification. To test their approach, 260 status updates posted within a week during the Tunisian revolution were collected in phase one. In phase two, three different sets of lexicons were created: emoticons or smiley faces, acronyms such as “gr8” that means great and “lol” that means “laugh out loud,” and interjections such as “haha” and “Wow” were used. Approximately 30 different lexemes were used. In the third phase, pre-processing included removing stop words that do not affect sentiments since they are neutral words, and stemming was used to enhance system performance. Moreover, the roots of opinionated words were used. POS and n-grams were used as features in phase 4. In the last phase, an updated status is classified as positive or negative. The highest achieved accuracy was 75.31% using SVM outperforming NB, which achieved 74.05%.

Hamouda and El-Taher (2013) used different ML techniques to classify sentiments present in Facebook comments. Although their work is similar to ours in terms of nature and source of data, the main difference between our work and theirs is that they study relative polarity of a comment, i.e., whether a comment is for or against the main post, whereas our work classifies comments in general for having positive, negative, dual, spam, or neutral sentiments. Another difference is that we are using custom-made LB classifiers instead of ordinary ML classifiers, although we do use ML classifiers for baseline results. Finally, the authors use three different classes (positive, negative, and neutral) whereas we use five. Their approach consists of three main phases: pre-processing, feature selection and classification. During pre-processing, stop words and long comments (more than 150 words) are removed. In phase two, nine different features are selected, most of which are counters of words such as number of words of comments, common words, and counter of negating words in comments and their comments; all of the features were normalized to have a value between 0 and 1 and vectors representing comments are then created. In the classification process, the target is to classify a comment as agree, disagree or neutral with respect to the main post. For this purpose, 2400 comments,

collected from 220 posts, were used; the comments were equally distributed as agree, disagree and neutral. Comments were manually classified to train three different classifiers: NB, SVM and decision trees. The highest accuracy reported was 73.4% when SVM was used.

In our work, we try to provide additional resources for Arabic SA, implement an LB classifier, compare the efficiency of LB classifier compared to ML classifiers, and study the effect of using NLP tools, such as Named Entity Recognition (NER) and POS tagging, on classification accuracy.

2.3.2 Arabic NLP

The Arabic language is a morphologically complex language (Habash et al., 2005). It has comparatively fewer and weaker tools and resources compared to English. Arabic NLP (ANLP) applications attracted many researchers in the last two decades given that it is one of the official UN languages and is spoken by hundreds of millions around the world. Many NLP applications such as machine translation, question answering, recommendation systems, sentiment analysis and others require variety of computational linguistic tools and datasets.

The Arabic language provides a clear case of diglossia (Ferguson, 1959), where more than one version of the same language are used at the same time: Classical Arabic is the language of the Holy Quran read by Muslims, MSA is used in formal communication and by scholars, and the IA is used in informal communication. This diversity makes it hard to create one tool or resource that can cope with the differences among the different versions. IA alone has many forms depending on geographical locations among countries and within the same country (Levantine, Egyptian, Gulf, etc.) The hardest form of Arabic is the IA because it does not have a specific grammar yet, neither has it adhered to spelling rules, especially when used in writing for social media. Among the issues that make the Arabic language a complex one is the lack of capitalization, the rich morphology, and the use of diacritics. Diacritics are short vowels that fully change the meaning of a word.

Habash et al. (2005) suggested specifying the features of the dialect to make it closer to MSA and then applying MSA NLP tools. Another approach proposed by Farghaly (2004) was to create an inter grammar that includes all the common core rules among all three versions of Arabic language.

The optimal solution seems to create separate resources and tools that consider that nature of each version. Farghaly and Shaalan (2014) and Habash (2010) studied ANLP and discussed the

challenges it faces. They specifically addressed the morphological complexity of the Arabic language and its effect on different applications along with proposed solutions.

Concerning the corpora needed by different NLP applications, there has been a significant increase in the number of available corpora. Despite their small number compared to English, there are available corpora for all versions of Arabic such as the MSA corpus introduced by El-Hajj and Koulali (2013) and the Informal Arabic corpus described by Itani et al. (2017b) and many others. Zaghouni (2017) conducted a survey about the major Arabic corpora currently available along with their characteristics and usages.

Concerning the tools, the major ones needed by ANLP applications are POS taggers, morphological analysers, stemmers, NER systems, tokenizers, and automatic diacritizations. Pasha et al. (2014) provide an Arabic language analyser, MADAMIRA, which has many tools within it such as POS tagging, tokenizing, and stemming. Green and Manning (2010) describe Stanford's statistical parser that can be used for Arabic language. Word segmentation systems were developed by Monroe et al. (2014) and Abdelali et al. (2016). Another important resource for Arabic POS tagging is Stanford's POS tagger addressed by Toutanova et al. (2003).

2.3.3 Arabic Sentiment Mining Literature

This section addresses main research works related to Arabic sentiment analysis with an emphasis at the end on those focusing on social media. This work is similar to some of the approaches mentioned below in terms of the pre-processing followed, usage of sentiment lexicon, addressing negation, and using NLP tools. However, it differs in terms of number of classes used, the analysis it offers for the incorrectly classified comments, and the diversity of dialects and lexicon, which reduces the efficiency of dialect-specific tools (such as MADAMIRA's Egyptian dialect NER and POS tagger).

Mining opinions from social media were first described by Abbasi et al. (2008) by applying SA to web fora, which have comments written in both English and Arabic. The approach employs syntactic features (Word/POS tag n-grams, phrase patterns, punctuation, etc.), semantic features (Polarity tags, appraisal groups, and semantic orientation), link-based features (Web links, send/reply patterns, and document citations) and stylistic features like vocabulary richness, special characters frequencies, and structure of words. A new algorithm named Entropy Weighted Genetic Algorithm (EWGA) was developed for enhanced feature selection. EWGA reduced the number of features to be used from more than 12000 into 500. Finally, an SVM

classifier was used and the highest accuracy achieved for the Arabic text was about 93%, which is impressive as it is similar to high inter annotations agreement reached by humans. The approach was applied independently on MSA and English.

Farra et al. (2010) present two approaches for sentence-level sentiment mining and one approach for document-level. Two main problems were addressed: categorize sentiment of a sentence with different POS as positive, negative or neutral and categorize the dominant sentiment of a document containing many opinions given the classes of sentences present in document.

The first sentence-level sentiment mining approach relies on grammatical nature of the Arabic language. The approach was tested on 29 sentences extracted from English movie reviews and translated to Arabic. SVM classifier achieved an accuracy of 89.3% using 10-fold cross validation for the training set.

The second sentence-level sentiment mining approach relies on syntactic and semantic features. A decision tree classifier was used in two different modes: (1) features were given their sentiment manually and this resulted in 80% accuracy and (2) features were given their sentiment by referring to the dictionary and this resulted in 62% accuracy. The authors claim that the low accuracy obtained when using the dictionary is because the sentiment of words is context-dependent.

The authors then addressed document-level sentiment mining by using the two approaches mentioned above to classify a document using the known sentiment of the sentences of the document as input to the classifier after dividing the document into chunks (number of sentences). The highest accuracy (87%) was obtained for four chunks and after excluding neutral documents. The document is classified based on semantic contributions of chunks.

El-Halees's (2011) results showed that using one classifier for document-level SA gave poor results and three classifiers were used sequentially to increase performance. Consecutive use of classifiers increased accuracy from ~50% when one classifier was used to ~60% when two classifiers were used to ~80% when three classifiers were used. All classifiers gave better results when classifying positive documents because the negation (polarity inverters) present in negative documents increases the complexity of the classification process. No solution to this problem was reported in the paper.

Abdul-Mageed et al. (2011) investigated the subjectivity and SA of MSA. Three experiments with different pre-processing setups were run on annotated corpus containing news documents. Language-independent and Arabic-specific morphological features were used. The authors proved that language-dependent features and domain-related polarity lexica increase performance. Native speakers annotated 400 documents containing 2855 sentences. Each sentence was labelled as objective (OBJ), subjective-negative (S-Neg), subjective-positive (S-Pos) and subjective-neutral (S-Neut). A manually-created polarity lexicon containing 3982 news-related adjectives was used. The language-independent features included adjective, n-grams and unique (words that occur less than five times). The approach consists of two phases: (1) A binary classifier is used to classify sentences as objective or subjective, (2) SVM binary classifier is used to classify subjective sentences as S-Neg or S-Pos (S-Neut was disregarded). The authors reported that adding morphological features improved classification accuracy by 0.15% in case of subjectivity and 1% in case of sentiment. The increase in performance after using language-dependent features proves the authors' claim that classification performance improves when using language-independent and language-specific features.

Abdul-Mageed et al. (2014) addressed subjectivity and SA at the sentence level of morphologically rich languages (such as Arabic). The authors designed a system called SAMAR, which operates on Arabic textual data of social media. The system handles four main objectives: (1) Arabic SSA of morphologically rich languages, (2) Feasibility of using standard features for SSA for social media given the small size of comments usually used in these media, (3) Effect of different dialects and (4) Social Media-specific features. Abdul-Mageed et al. (2011) discussed the first objective in detail, where the authors showed that by considering the morphological complexity, the classification performance increases. Since SSA is highly dependent on lexicons, systems used for SSA for English cannot be directly applied to Arabic because of the complexity of Arabic. Different lemmatization setups were tried and performance varied accordingly, which supports the authors' claim about the effect of morphological complexity on classification performance. Classification using SVM is done in two stages. In stage one, a sentence is classified as objective or subjective, and in stage two the subjective sentiments are classified as positive or negative (neutral and mixed classes are disregarded). Different types of features were used: morphological features (word forms and POS tagging), standard (Unique, when a word

occurs in low frequency and polarity lexicon), dialectal features (MSA or IA) and genre-specific (user ID, gender and document ID).

Rushdi-Saleh et al. (2011) presented an Opinion Corpus for Arabic (OCA). The authors reported the three main difficulties faced when generating the corpus:

- Unrelated Comments: They are irrelevant comments posted on blogs and different from the discussed topic. Such as when users start chatting or discussing different topics.
- Transliteration: This is a common phenomenon in online comments; authors use Roman letters to write in Arabic. Those who know Arabic and English understand the meaning. Such cases do not follow spelling rules and many possible variants are possible.
- Using Foreign Languages: Authors may use English, French, or other languages to comment in Arabic sites.

A total of 500 movie reviews were collected from different sites, half of which are positive and the other half are negative. The reviews were processed by removing HTML tags and special characters and manually correcting the spelling mistakes. Afterwards, each review was tokenized and stopwords were removed, then stemming was applied. The sites selected were those using MSA. Three main issues were noticed concerning rating of reviews:

- Rating System: Different sites use different scales for rating a movie: some use a scale of five, some use a scale of 10, and others use binary rating: good or bad. In numeric scales, movies with review above average were considered good.
- Effect of Politics and Religion: Comments can be affected by the commenter's background in politics and religion regardless of the artistic criteria.
- Proper Nouns: Movie and actor names are translated into Arabic in some cases and kept in English in others.

Two different classifiers were used to evaluate the corpus: NB and SVM. SVM achieved better results. The highest F1-measure computed was 0.9 when n-grams were employed. Zaidan and Callison-Burchm (2011) presented a dataset of dialectal Arabic comments extracted from three news websites. The comments were manually classified according to their dialect (Egyptian, Gulf, and Levantine). They also presented a system that can automatically detect the dialect of a comment. Such a system is one step towards converting IA into MSA, which will enable

researchers to use tools available for MSA. The dataset contains 1.4M comments having 52.1M words. Language modelling was used to classify comments as written in MSA or IA, and to classify the dialect being used among a set of dialects. The authors reported 77.8% accuracy for the first setup (MSA vs. IA) and 83.5% in the second (detecting which dialect is being used). Almas and Ahmad (2007) studied SA of financial news written in Arabic and Urdu. Their approach classifies sentences as positive, negative, or dual. Randomly selected 30 Reuters Arabic and 20 Reuters English–UK documents were used to evaluate the approach. Two human taggers, one worked on the English version of the documents and the other worked on the Arabic and Urdu version, classified the documents as positive, negative, dual, or neutral. “Unknown” was used to label documents of unknown sentiment. The English tagger classified the documents according to what is negative or positive to the English economy, the other according to what is negative or positive to Middle Eastern economies. Their approach uses Quirkian notion, in which a linguistic unit is related to the frequency of occurrence of that unit. A local grammar is then developed, which considers the significance of lexemes in special and general corpora, i.e., how important a lexeme can be in a specific domain. The authors reported 28.8% accuracy for the Arabic corpus and 20.1% accuracy for the English version.

Itani et al. (2012) used Facebook comments written in IA as corpora for SA. Classifying sentiment was done based on searching for lexemes that are commonly used to express negative opinions, positive opinions or spam. Different sets of lexemes were created during the manual classification of the corpora and these sets were used as references to classify comments. Five different classes were used in the classification (negative, positive, neutral, dual and spam). Different setups were conducted, where a setup specifies which set of lexemes to use in the classification process, and the highest recall and precision reported were 50% and 85% respectively.

Al-Kabi et al. (2016) suggested a prototype to build a corpus for SA of MSA. Their corpus consists of 250 topics distributed equally among five domains: Economy, Food-Life style, Religion, Sport, and Technology and collected from The Maktoob Yahoo! website. The corpus has 1296 reviews associated to it, and the authors have provided different statistics such as number of words per review, per topic, and per domain. Also provided is the percentage of each dialect among the reviews with MSA (65%) and Egyptian (15%) being the two major dialects.

The authors have also annotated the corpus per gender of reviewer and sentiment of review using five different classes: negative, positive, neutral, spam (or irrelevant), and unknown. The authors claim that the majority of Maktoob users prefer to use MSA to enable other Arabs to understand their reviews.

Adouane and Johansson (2016) provided linguistic resources for Gulf Arabic sentiment analysis. The authors collected 4072 restaurant reviews; the reviews were negative, positive, neutral, or mixed. However, for the classification setups, only negative and positive reviews were used. Four different setups were conducted using an NB classifier; each setup used a specific lexicon or a combination of lexicons. The highest accuracy reported by the authors is 90.54% and was achieved when using the Gulf Lexicon alone. It is worth mentioning that negation was addressed by reversing the sentiment of the opinionated lexeme whenever directly preceded by an inverter.

El-Beltagy (2016) provided word and phrase level sentiment lexicon for MSA and Egyptian dialect. The lexicon consists of 5953 entries, 55% of which are in MSA and 45% are Egyptian. The author collected many of the entries from her social media posting. The lexicon was tested on two twitter datasets, one Saudi and the other Egyptian. The highest F1-measure achieved was 89.7% for binary classification (positive or negative) and 71% for three-class classification (positive, negative, and neutral).

Zaghouani (2017) conducted a survey about the freely available Arabic corpora. The aim of the survey is to boost the availability and easy access to Arabic corpora that are considered scarce compared to what is available for the English language. The availability of these corpora is essential for advancements in Arabic NLP application. The author divided the corpora into 6 categories: Raw Text Corpora, Annotated Corpora, Lexicon, Speech Corpora, Handwriting Recognition Corpora, and Miscellaneous Corpora types. Each category included several sub-categories. The survey included 66 corpora, and for each of them the author mentioned the creators, the name of the corpus, and the size. The survey does not include sentiment corpora.

Al-Ayyoub et al. (2017) addressed Aspect-Based Sentiment Analysis (ABSA) with a focus on Arabic language. Specifically, their work focused on Arabic Laptop Reviews and their research demonstrates how a dataset for the reviews was constructed. Their approach is in line with SemEval16-Task 5 annotation scheme (Task 5 is dedicated to ABSA). The annotation addressed two issues: predicting the aspect category and its sentiment polarity class, both applied on two

levels: sentence-level, and review-level. Given an opinionated review (or sentence) about an entity, the aim is to determine all targets and the polarity of opinions towards them (using three labels: positive, negative, or neutral). The authors demonstrated how the dataset was constructed and how an SVM classifier was used to classify it. Results reported show high accuracy in sentiment classification (F1-measure = 0.732) and low accuracy in predicting the aspects (F1-measure = 0.315).

Salameh et al. (2015) studied the effect of translation on sentiment analysis. Specifically, they addressed the sentiment of Arabic social media posts translated to English. The authors discussed three methods used to classify the sentiment of non-English texts: (1) using a language-specific sentiment analysis system, (2) using English sentiment analysis on manual translation of the source language texts, and (3) using English sentiment analysis on automatic translation of the source language texts. The authors worked on five datasets, and the result showed that using English sentiment analysis system on translated texts does not dramatically degrade performance, with manual translation outperforming automatic translation.

Alwakid et al. (2017) investigated the challenges that face Arabic SA and provided an approach that starts by linguistic pre-processing that addresses the complexity of format of Arabic tweets with a focus on Arabic dialects. NB and SVM classifiers were used to classify sentiments of tweets. Their work also suggested a framework for implementing domain-specific and knowledge-assisted sentiment classification.

This study adopts two main approaches mentioned in literature, namely the LB and NB classifiers and compares their results in chapter 6. Moreover, the corpora provided by this research were classified using some of lexicons used in literature in addition to using the lexicon provided by this research to classify some of the corpora mentioned in literature. Details about corpora and lexicons used are found in chapter 6.

2.3.3.1 SemEval-2017

One major on-going cycle of development is the International Workshop on Semantic Evaluation (SemEval) that has different subtasks related to semantic analysis. The works summarized below try different approaches to address the subtasks of the workshop.

SemEval-2017 Task 4: Sentiment Analysis (Rosenthal et al., 2017) has two new changes than previous year, namely: introducing Arabic for all subtasks and making information from the profiles of Twitter users, who posted the target tweets, public. The task Sentiment Analysis in Twitter started in 2013 (Wilson et al., 2013; Rosenthal et al., 2015; Nakov et al., 2016a; Nakov et al., 2016b). In 2015, the task started addressing sentiment towards a topic, and in 2016 it included tweet quantification and five-point classification (highly positive, positive, neutral, negative, and highly negative) to be similar to the rate used by major corporations such as Amazon, Yelp, and TripAdvisor. Task 4, which was addressed by different teams, included four subtasks:

(A): Classify a tweet as positive, negative, or neutral.

(B): Given a topic and a tweet, classify the sentiment expressed in the tweet towards the topic as negative or positive (2-point scale)

(C): Given a topic and a tweet, classify the sentiment expressed in the tweet towards the topic using 5-point scale (highly positive, positive, neutral, negative, and highly negative).

(D): Given a set of tweets about a topic, cluster negative and positive tweets

(E): Given a set of tweets about a topic, study the distribution of tweets among the 5-point classes.

The authors used different classifiers such as NB, Maximum Entropy, and Random Forest.

The highest F1-measure achieved for subtask A was 0.61 (El-Beltagy et al., 2017).

El-Beltagy et al. (2017) described two systems that were used in three subtasks of SemEval-2017 Task 4, namely, subtasks A, B, and D mentioned above. For subtask A, 13292 tweets were used for training and 671 were used for testing. An NB classifier that relied on weighted sentiment lexicon was used for classification and achieved an F1-measure of 0.61. Vectors including different features such as the number of positive and negative lexemes, presence of hyperlinks, and size of tweets were used to construct the input vectors. As for subtask B, three classifiers were used, and voting was done to label the tweets, F1-measure of the three classifiers ranged between 0.72 and 0.759. The output of subtask B was converted to fit the input needed by subtask D.

Htaït et al. (2017) addressed subtask A of SemEval-2017 Task 4, which is to classify a tweet as negative, positive, or neutral. Their approach uses a set of sentiment words and the sentiment relation between the seeds words and other words is determined using the cosine similarity among the word embedding representations. The initial set of sentiment words was taken from public annotated tweets. Their lexicon includes negative and positive terms only. The approach was tested using SemEval data. The authors reported an average F1-measure of 0.561 for English and 0.469 for Arabic.

Mulki et al. (2017) also worked on subtask A of SemEval 2017 task 4. Two approaches were proposed; one is supervised and uses SVM and NB classifiers, and the other is unsupervised and uses a lexicon-based classifier. Both approaches start by preprocessing the tweets and cleaning them from noisy data such as hashtags, dates, usernames, etc. The proposed models operated on a dataset consisting of 2684 labelled tweets for training, 671 tweets for tuning, and 6100 tweets for testing. The supervised models achieved an F1-measure of 0.416 and the lexicon-based model achieved an F1-measure of 0.342.

Baly et al. (2017a) demonstrate four systems that were implemented to address SemEval-2017 task 4, Opinion Mining for Arabic and More (OMAM) Systems. Concerning subtask A, they evaluated the English sentiment analysis methods on Arabic tweets, and for the rest of the subtasks, the authors used a topic-based approach to predict the domains or topics of tweets, and then use this knowledge to determine their sentiment. For subtask A, results show that using English methods has reached a threshold with no major improvement (average F1-measure is 0.422), and for the remaining subtasks, the following were observed:

- For subtask B, ignoring the topic achieves best performance.
- For subtask C, using a topic-specific sentiment classifiers, and supporting them with domain-specific sentiment classifiers, achieved the highest performance for subtask C.

Baly et al. (2017b) addressed the main challenges facing Arabic SA in Twitter. They introduce a characterization analysis of tweets from diverse Arab regions to show how Twitter usage varies across the regions. They also study how specific tokens such as mentions, pictures, hashtags, and URLs may contain subjective information that can affect the tweet's sentiment. The authors compare the performance of two different models used in opinion mining: one that uses feature engineering and another that relies on deep learning. The first model used semantic, syntactic,

and surface features and a SVM classifier, whereas the second model used Recursive Neural Tensor Networks (RNTN) (Socher et al., 2013). The classifiers ran on 3315 tweets that belong to three classes (negative, positive, and neutral) and RNTN achieved an average F1-measure of 53.6% compared to 43.4% achieved by SVM.

2.4 Negation Literature

Polanyi and Zaenen (2006) introduced the idea of valence shifters. A valence shifter is a word that intensifies (such as “so” in “so strong”), weakens (such as “slightly” in “slightly hard”) or flips (such as “not” in “not easy”) the polarity of a sentimental word. Valence shifters used in the English language can be divided into two main categories: Sentence-based and discourse-based.

Sentence-based valence shifters are:

- Negatives and intensifiers: words such as not, never, none, no one, neither, etc., which can flip the polarity of a term from negative to positive as in “not bad at all” and from positive to negative as in “not good at all”. Besides inverting polarity, some words can intensify the sentiment such as “badly” in “badly injured.” Others can weaken the sentiment such as “slightly” in “slightly interested”.
- Modals: may be used to assume future consequences that are built on the probability of an event to happen. Opinionated words within the range of the modal will not behave normally. For example, in the sentence “If Marwan were lazy, he would fail in his exams” there are two negative words, “lazy” and “fail;” however, the sentiment of these words is affected by “would,” and we can understand from the sentence that Marwan is neither lazy nor did he fail his exams.
- Presuppositional Items: are words that can shift the valence of words because an event did not meet expectations such as the word “almost” in “he almost passed” means that he did not pass, the same thing can be said about “barely” in “the water was barely enough.” In these two examples, the presuppositional items shifted the neutral and positive sentiments into a negative sentiment or leaning toward negative. Different parts of speech have the same effect such as “failed” in “failed to pass” and “impossible” in “impossible to enjoy.”
- Irony: intense positive or negative words may express opposite polarity such as “genius” in “the genius professor did not know how to solve an easy problem”. Although “genius” is a positive word, the way it was used in context gives the exact opposite meaning.

In this work, we address the first category only, negatives (inverters will be used hereafter) and analyse their effect on SA. The main reason for choosing the first category is that inverters are explicitly specified in Arabic grammar. Although other sentiment shifters exist, there is no specific categorization of these sentiment shifters. Such set of words have not been collected and analysed before but will be part of our future work.

Discourse-based valence shifters are:

- **Conjunctions or Connectors:** words such as but, although, however, etc. can affect opinionated words within their range. Consider the word “mean” in “Although he is mean, he treats animals well”, “mean” is a negative word, but since it was used after “Although”, its negativity was neutralized by the positive second phrase of the sentence.
- **Discourse structure:** Sentences may consist of a dominant part and an illustrative part that supports the dominant part. If the dominant part was opinionated, then the illustrative part will intensify the sentiment even if it was neutral by itself. Consider the sentence “He is a great fisherman. He caught 5 kilos of fish yesterday.” The first sentence is positive due to the presence of the word “great” and the second sentence is neutral. However, the position of the neutral objective sentence directly after the opinionated sentence intensifies the polarity present in the first sentence by providing facts that support it.
- **Multi-entity evaluation:** If textual data contain many positive words about different objects and many negative words about one object, counting the number of negative and positive words to classify the text would be misleading because although many negative words have been used to criticize one object, many other objects were positively commented on, which means that total sentiment should not be negative.
- **Reported speech:** reporting sentimental text does not imply that it is accepted by the author, the sentence “he said the movie is great” does not mean that the user agrees, and thus the sentence cannot be considered as positive. However, in the sentence “He said the movie is great, and I totally agree,” the second phrase supports the first one resulting in a totally positive sentiment. Using the same argument, we can say that the sentence “he said the movie is great, but I don’t agree at all” is negative since the word “great” was neutralized by the second phrase.

- Subtopics: Long documents may be split into subtopics, each having its own sentiment. In such cases, there may not be a specific sentiment for the document as a whole; the sentiment is relative depending on the subtopics covered.

- Genre constraint: Topics like movies or books reviews contain information about the book or the movie themselves and about events inside them. For example, a review such as “It was a great movie. It tells about a mean person who lived a miserable life suffering from poverty and injustice” should be considered positive even if the number of negative words is bigger than the number of positive ones because the purpose of the review is to classify the movie as good or bad regardless of what the movie is about.

Although no quantitative analysis for the effect of these valence shifters was given by Polanyi and Zaenen (2006), nor any approach was suggested to efficiently employ them in sentiment classification, the work is significant in highlighting which features may be useful in inverting/shifting sentiment of words. A specific list of each type could be prepared and used in an opinion classification system.

Three main contributions related to the effect of negation were presented by Jia et al. (2009). The first is an approach named SCT, which is used to determine the scope of negation when an inverter (such as none, not, never, or barely) is present in a sentence. The second contribution is a method to determine the polarity of a segment of a sentence containing an inverter. The third is a study of the effect of introducing the concept of scope of negation on opinion retrieval system.

Abbasi et al. (2008), and Pang and Lee (2004) resolved negation by using an LB approach. They used a list of negating terms to identify negations and shifted polarity according to whether a term from the list existed or not. In their work, opinionated words were given signed weights. For example, -5 indicated extremely negative and +5 indicated extremely positive. Negation would then shift the polarity by decreasing/increasing the signed weight by a fixed amount equal to 4. So a negation would not totally invert the weight of a sentiment for example from 3 to -3 or from -5 to 5. For example, if “awesome” had a weight of +5, then the weight of “not awesome” would not be -5, but rather the subtraction of the fixed amount (4) from the weight (+5) to obtain a 1. This meant that “not awesome” still had a positive sentiment that is less than “awesome.” Maynard and Funk (2011) reported that the presence of negation increases the complexity of classification. The authors did not specify how complexity increases nor did they suggest a

solution. Experiments were conducted on 1143 documents (635 positive and 508 negative) of three domains (education, politics and sports), the authors reported that classification worked better with positive documents compared to classifying negative documents. The authors linked the worse performance in negative documents to negation but were not clear on how they reached this conclusion.

In the works of Hamouda and El-Taher (2013) and Abdul-Mageed et al. (2011), the frequency of inverters in a comment was used as one of the features in the classification. However, the behaviour of inverters and their exact effect on the classification process were not addressed. Hamouda and Akaichi (2013) assumed that negating terms always precede the targets directly and hence targets' polarity was inverted whenever preceded by an inverter. A similar behaviour of negating terms was assumed by Hamouda and El-Taher (2013), and negating terms were used as classification feature. This fact is related to the nature of the language, which is also true for English in that inverters precede their target almost always. Improvement in classification performance after considering negation supports this claim. Our statistics concerning scope of inverters are in harmony with this assumption as expected according to grammatical rules of Arabic language (Abdul-Mageed et al., 2012).

Negation's effect on SA in Arabic was also mentioned by El-Halees (2011). The author reported an increase of performance of classification by ~1% when negation was considered. However, no details about the behaviour of different negating terms were presented. El-Beltagy (2016) mentioned that although the presence of an inverter may flip the polarity of the opinionated lexeme following it, there are odd cases where the presence of inverters may affirm the polarity of the lexeme following it.

In our approach, all inverters were treated as if they have the same behaviour, i.e. they flip polarity of target. The noisy nature (when having different meaning or appearing as parts of other words) of inverters was not discussed. This work addresses the effect of negation and proposes a straightforward potential solution to help improve classification performance in a number of cases.

Summary

Chapter 2 starts by providing context background and briefly visits NLP related paradigms such as IR, data mining, classification algorithms, etc. Afterwards, the chapter covers the main

researches in SA literature and specifically in social media. Then the challenges facing Arabic NLP are addressed before highlighting the main approaches and findings related to Arabic SA. The chapter also focused on SemEval-2017 due to the relevance of its subtasks to this research. The chapter ends by covering major works related to effect of negation. The results reported, especially those of SemEval-2017, show that there is room for improvement, and that Arabic literature can benefit from additional annotated resources. Authors of the researches mentioned earlier highlighted many issues that hinder Arabic SA:

- 1-Limited number of corpora annotated for SA purpose.
- 2- Limited number of opinionated lexicons that can be used to classify IA text.
- 3-Lack of deep analysis of reasons behind incorrect sentiment classification.
- 4-Negation can degrade classification performance, and therefore resolving its effect can improve classification results.
- 5- Social media have massive amount of IA comments. Classifying sentiment of these comments can be beneficial to decision makers.

This work addresses the limitations mentioned above and highlights areas that need further research in the future. The third point, in specific, sheds light on issues that needs additional resolution to enhance classification performance.

CHAPTER 3: Research Philosophy

A core assumption in sentiment analysis is that people tend to express their emotions and feelings on social media platform using expressions similar to those that they use in real life. In other words, we assume that words of expression exist independently of the platform used; this assumption will place our research under Positivism. We also mainly adopt deductive and quantitative research approaches as explained in the upcoming sections. However, the work does also take account of the fact that sentiments may be expressed differently in different domains. Hence, in chapter 4 the techniques developed were applied in different domains to explore their domain-independence.

3.1 Introduction

Chapter 3 discusses the research methods followed at different stages. It argues why each methodology was used and how it serves the research objectives and helps in answering the research questions. The chapter also briefly introduces how data collection and analysis were done. Finally, the chapter addresses ethical issues associated with social media data and research data management. For the reader's convenience, the research aims and questions that were mentioned in chapter 1 are stated below again:

Research Objectives:

- Investigate (identify) classical techniques used in sentiments analysis (SA) with focus on Arabic language.
- Develop a better understanding of SA by developing an LB classifier that uses a sentiment lexicon to classify SM comments written in IA according to their sentiment.
 - Construct an annotated a corpus (large collection of text) to be used for SA.
 - Construct an opinionated lexicon (a dictionary that assigns a polarity (positive, negative, etc.) to words instead of meaning)
- Identify the main reasons behind incorrect sentiment classifications
- Make recommendations concerning improving the system and method of classification until it reaches saturation level at which the classification results become similar to agreement results of human classifiers.

Research Questions:

1. How can we get a better understanding of SA of SM comments written in IA?
2. How can we improve sentiment classification of SM comments?
3. What are the main reasons behind incorrect classification of IA when LB classifier is used?

3.2 Types of Research Approaches

Our choice of research methodologies whether during data collection or analysis was governed by the above research objectives and questions. These belong to the discipline and practices computational linguistics since it involves extracting and classifying opinions (Keshtkar, 2011). Following that practice, a deductive and qualitative approach was adopted. We now briefly compare two main categories of research methodologies and highlight how our choice matches the research's scope.

3.2.1 Deductive vs. Inductive Approach

Research methodologies can be divided into two categories in order to reach research objectives. The two categories are quantitative versus qualitative and inductive versus deductive research approaches.

Briefly, deductive research can be looked at as a top down approach that starts from general knowledge and zooms into more specific one (Dudovskiy, 2017). It is driven by a theory of the subject being addressed and research questions regarding that theory. The approach to answering the questions thus involves narrowing down the specifics of the questions with aim of validating them and in doing so contributing insight to the body of knowledge of the subject.

On the other hand, inductive research approach is considered a bottom up approach (Bradford, 2015). It begins with specific observations then aims to expand upon them to identify hypotheses and theories relevant to the observations and insights regarding the domain being examined. The approach starts at a pattern or observation then tries to formulate a hypothesis or put the observations in a theoretical frame.

This dissertation will follow a deductive approach because it starts from the theory that states that the sentiment, or domain, of a text can be governed by presence of specific words. An LB

classifier uses this theory and attempts to classify a given text according to predefined labels. In this work, we try to apply this concept on IA comments extracted from social media.

3.2.2 Qualitative vs. Quantitative

Research techniques can be divided into qualitative or quantitative techniques. A qualitative research deals with non-numerical data focusing upon text and audio (even pictures and video) as primary data. One simple example is the use of open-ended questions in a questionnaire. The answers to such question have to be interpreted and understood by the researcher. Qualitative methods can also be considered as explanatory methods trying to find the how and why related to the subject being addressed. Hence, it tends to ask broad questions and collects information from participants or phenomena via observations, surveys, questionnaires, etc.

The second class of methodology is quantitative research that is based on numerical data. The goal of quantitative research is to implement and exploit mathematical models using existing theories and hypotheses related to some phenomenon. The concept of measurement is core to quantitative research since it provides the main link between theoretical observation and mathematical values.

In this work, the quantitative approach is mainly followed. SA is traditionally quantitative by nature because the underlying emphasis is upon automating the classification process and optimizing the performance of the classifier. However, there is a qualitative aspect when considering the manual classification of comments as well when categories of incorrectly classified comments are being analysed and discussed. In these cases, the research relies upon knowledge of the IA and consensus between the interpretations of native speakers.

3.3 Data Collection Technique

Data was collected in two different ways:

1-Copying comments: Public FB comments on two public pages were manually collected by copying the comments, 1000 comments from each of the pages mentioned next. The processing of the copied data is described in chapter 4. One of the pages is related to news (<http://WWW.facebook.com/AlArabiya>) and the other is an arts page (<http://WWW.facebook.com/MBCTheVoice>) that covers news related to singing competition. In both cases, the data that has been collected consists of the comments that users post in respond to posts posted by the pages' owners.

2-Requesting input from Facebook users: In addition to using the corpus annotation to extract opinionated lexemes, and in order to boost the lexicon, 100 Facebook users were asked to provide positive and negative domain-independent words or phrases that they would use to express a negative or positive opinion, they provided a total of 541 lexemes. All phrases and words are written in IA and are less than a sentence long. Statistical details will follow in chapter 4. The 100 users are the author's FB friends of different Arab nationalities but mainly Lebanese.

3.4 Data Analysis

Data analysis took place at three different levels:

Corpus annotation: A corpus is a collection of data, although it is usually textual, other types such as audio do exist. A corpus usually has a theme such as corpus of newspaper articles or corpus of all grade 5 math exams, etc. Annotating a corpus means giving a label to each record in it. For instance, annotating a corpus consisting of 1000 articles means giving each article a label from a predefined set of labels such as science, sports, arts, etc. These labels are used to train classifiers and test their performance.

In this work, the corpus consists of 2000 FB comments written in IA. Following data collection, expert native speakers of Arabic labelled each comment as negative, positive, dual, neutral, or spam. Annotation rules will be detailed in next chapter. This annotation depended on existence of words or phrases, which according to the human tagger, were behind giving the comment its sentiment.

Constructing the lexicon: In addition to giving a label to the comment itself, the opinionated words mentioned earlier were put in different sets according to their sentiment. Three main sets were created for this purpose: negative, positive, and spam. These three sets constitute the sentiment lexicon.

Classification performance and analysis of incorrectly classified comments: Following constructing the corpus and the lexicon, a classifier was implemented and its classification results were compared against manual classification done by native speakers. Afterwards, the incorrectly classified comments were analysed to study the main reasons behind incorrect classification, this was also done based on quantitative measures with optimising the classifier's performance being the main target.

The performance of the classifier was measured using the classical F1-measure that equally uses precision and recall. There exist other F measures that give different weights to precision and recall but were not used in SA literature. F1-measure is the one usually used since it gives equal significance to precision and recall. In NLP context, precision represents the fraction of retrieved documents that are relevant to the query whereas recall represents the fraction of the relevant documents that are successfully retrieved. A comment was considered to be correctly classified if there is an exact match between manual classification and automatic classification.

Briefly, the implemented classifier adopts a bag of words approach, addressed negation cases where inverters directly preceded their targets, and gave highest priority to spam lexemes, i.e., their presence would dominate the sentiment class. For each record to be classified, the classifier searches for spam lexemes, if found, then any other negative or positive lexeme is ignored. The classifier would search for the presence of negative and positive lexemes and check if they were preceded by inverters to flip the lexemes' sentiment accordingly. Finally, a record is classified as positive if it contained positive lexemes only, negative if it contained negative lexemes only, dual if it contain negative and positive lexemes and neutral if it does not contain no opinionated lexeme.

3.4.1 Research Approach

Following initial implementation of the classifier and analysis of results, the approach was assessed and then improved to become closer to the human classification of comments. The ultimate objective of the classifier is not to reach 100% accuracy as this is not possible even among humans and there exists a disagreement margin in classifying post; the objective, however, is to improve classification accuracy to an extent close to an acceptable inter annotation agreement level that may exist if group of humans were to classify a corpus, different researches marked this level in 80s and 90s (Somasundaran et al., 2008; Abdul-Mageed & Diab, 2011).

3.5 Ethical Considerations

In addition to ensuring that a research is done legally, a research needs to ensure that it has been done ethically. Ethical guidelines may vary in scope and phrasing, but they all agree on the core. The major points that ethics focus on are the rights of human dignity and safety, maximizing benefits, minimizing risks, respect for people, and justice. Although a research study may be

conducted legally and respecting a platform's terms and condition, a researcher needs to keep in mind that some users are more vulnerable than others due to many factors such as medical issues, educational background, etc., and therefore, should put more efforts to ensure the wellbeing of such users.

This section covers the ethical frame that governed the research. It starts by briefly defining intellectual property and copyright, explains the potential copyrights infringements that may take place on social media platforms, and provides a guideline that helps users avoid any unethical conduct. Afterwards, the chapter focuses upon Facebook's privacy settings and terms and conditions perspective of copyrights and when they can be claimed. It ends by describing how data collection was done ethically without violating users' copyrights or the FB's terms and conditions. Sections 3.5.1, 3.5.2, and 3.5.3 provide a zoomed out picture of what researchers using online data should consider during data collection and usage.

3.5.1 Intellectual Property

One starting point to enforce ethical approach is to respect intellectual property (IP). IP refers to works create of mind such as artistic, scientific, and literary work (“What is Intellectual Property”, 2017). According to World Intellectual Property Organization (WIPO), IP can be divided into three categories:

- Industrial Property includes patents for industrial designs, trademarks designs, inventions, and geographical indications.
- Copyright covers literary works (such as novels, poems, etc.), films, music, artistic works (such as drawings, photographs, sculptures, etc.) and architectural design.
- Rights scope include related performing artists and their performances, producers of phonograms and their recordings, and broadcasters and their programs.

3.5.2 Copyright

IP does not prevent researchers from using resources created by others, it simply clarifies what can be considered as an intellectually property. The second step is to use IP of others while respecting their copyrights. A copyright is a legal definition that describes the rights of creators of artistic and artistic works (“Copyright”, 2017). According to WIPO, the creations that can be copyrighted range from music, painting, books, films, and sculpture, to computer programs, maps, advertisements, sculpture, and technical drawings. Creators of copyrighted work have the

exclusive right to do and authorize the actions below, detailed legal frames about what and how can works be copyrighted can be found at WIPO's website (WWW.wipo.int):

- Reproduce the copyrighted work in copies in any format.
- Derive works based on the copyrighted work.
- Distribute copies of the copyrighted work to the public by sale or any other delegation of ownership, or by lease, lending, or rental.

3.5.3 Copyrights and social media

Social media have revolutionized communication done between people among each other, and between people and organisations. The booming of social media, however, increased the risk of copyright infringement (“Copyright”, 2017). Social media provides user-friendly platforms that enable easy sharing of content and posting mechanisms that appear side step copyright issues. Although platforms specify their terms and conditions regarding copyrights, the significant number of users and the ease of sharing data without considering terms and conditions, emphasis the risk of unethical practice. Assume for instance that someone has posted a photo of her having coffee at a certain coffee shop with the logo clearly visible in the photo and post it on her Facebook page. Who then owns the photo? She may believe she owns it and has given a copy to Facebook. However, what permission has she granted Facebook? In addition, the owner of the coffee shop does not have the right to copy the photo and post on his FB page as if the photo is his, otherwise, copyright infringement will be taking place. The owner can, instead, share the photo using the functionality provided by the social media, a functionality that acknowledges who the original owner of published material. The owner of the photo needs to be aware that privacy settings of the post govern who can see the photo, i.e., if it is public, all users may see the photo. On the other hand, the coffee shop owner needs to know that he cannot save the photo and then use it as if it was his own, what he can do instead is to use the *share* functionality, or contact owner of the photo and asks if he can have and use a copy of the photo. Although such infringements are easy to detect when it comes to photos and videos, the border become vaguer when addressing textual data. For instance, someone posting short phrases on his page such as “I am happy today”, or “congratulations” cannot claim copyrights for such phrases. A copyright, by definition, protects works that have a minimum level of creativity, works such as poems, short stories, novels, etc. Short phrases do not qualify to be copyrighted, this ensured the comments

used in this research do not qualify to be copyrighted, and neither using them will endanger the author's emotional, financial, or physical wellbeing.

When posting on social media, it is better to follow a quick guideline that would ensure safe and ethical posting. It is recommended that one ask oneself the following prior to posting:

1- Who owns the material to be posted? Common types of copyright owners include:

- Author of a written text such as poems and stories
- Photographers
- Composer of a musical piece
- Videographers
- Publisher of published works
- Creator of art such painting or sculptures
- Institutions at which any of these authors if the work was created in connection with their institutions

2-How to get permission to post copyrighted content?

If the content to be posted qualifies to be copyrighted, the creator of the work should be contacted, otherwise, the content can be shared given the social media tools such as Facebook's "Share" tool that automatically shows the original owner of the content.

3-When is posting of the content considered a "fair use"?

Before answering this question, one needs to briefly know what fair use is:

Fair use allows using content without getting permission from owners. It depends on different factors such as:

- the type of usage: usually a fair use license is given for non-profit, educational, and personal usages
- the copyrighted content is published
- the size of the content used in relative to the size of the whole work
- the effect of the use on the market or value of the copyrighted content.

It is recommended to check with the content owner in order to be on the safe side.

4- What are the consequences of infringing copyright?

Usually, a copyright holder's first action towards infringement is to send a letter requesting to stop the infringement. The copyright holder can hold a lawsuit to get court order that enforces removing the infringing content and claiming compensations depending on nature and size of the infringement.

3.5.4 Facebook's Privacy settings

Each social media platform has its own terms and conditions to which each registered user must agree on prior to registration. Since this work is based on FB's comments, the focus will be on FB's terms and conditions.

Although all social media platform have their own terms and conditions, they all agree on the key legal and ethical principles. Although a regular online user rarely reads all terms and conditions, one needs to be aware of the basics that would protect his data from being used without his consent, even if this were done legally. On the other hand, a researcher has additional moral responsibility when it comes to using online data or conducting research on or about the internet because many online users are not aware of their vulnerability. Some platforms continuously provide user-friendly privacy tips to raise awareness among users. For instance, when commenting on a public post, a help screen describes to the user that others will be able to see the comment. Below is an excerpt from FB's terms and condition ("Statement of Rights and Responsibilities", 2015) that specifies the main rules that govern sharing content and highlights the privacy settings needed for one to be able to claim copyrights over his intellectual property:

"When you publish content or information using the Public setting, it means that you are allowing everyone, including people off of Facebook, to access and use that information, and to associate it with you" (i.e., your name and profile picture).

Appendix C has more details about Facebook's terms and conditions.

3.5.5 Data Management

A research may include, operate on, or produce confidential or sensitive data. In any of these cases, the researcher needs to ensure that ethics core principles are respected. In this research, a plan was setup that would explain the nature of the data, the way it was collected and used, and how it is to be shared and with others. The plan not only ensures a proper monitoring of who is

using the data and why it is used for, but it also follows needed technical measures to ensure that it is safely stored. Having such a plan at early stages help drawing clear frame for data collection and usage (see appendix D).

3.5.6 Data Collection and Ethical Issues

Data collection was done in an ethical manner ensuring no copyrights infringement took place. Although our usage of data most probably qualify to be fair use, the following milestones ensured ethical collection and use of data:

1. Our work operates on short textual phrases that do not qualify to be copyrighted due to their length, absence of creativity of any form, and their nature, i.e. dialectal phrasing not constituting a story, a poem, or any other literary work.
2. The collected comments were posted by users under the Public setting.
3. Although data collection and use were done ethically and legally, anonymization was applied to add another layer of security to users should they believe that tracking the posts back to them may in any way be harmful to them.
4. As a final layer of ensuring ethical approach, owner of the FB pages whose comments were used were contacted and their consent was taken.

3.5.7 Ethics Scrutiny

In addition to all that has been mentioned, there are individual and institutional responsibilities that add another layer of integrity to the work. Although it may be enough to ensure that research has been conducted legally and according to the terms and conditions of the platform used, the researcher's and institution's role is to enforce ethical scrutiny. Institutions usually channel the research through an ethical frame observed by ethics committee to ensure no violations took place. The checklist followed by institutions may vary but they usually provide a similar framework. In the case of this research, the main point tackled by the committee ensured that data collection and usage was done ethically and that data collection and usage will not cause any harm to any online user neither physically nor emotionally. This was applied by making sure that used material do not qualify to be copyrighted, terms and conditions have been respected, owners' consent have been taken, and authors of collected comments.

It is worth mentioning that FB users whose comments were used to construct the corpus did not know that their comments have been used. Although data collection was done legally, the nature

of data and the pre-processing done ensures guarantees the wellbeing of users and that they may not be harmed in any manner. Moreover, none of the comments used is a work of art that can be copyrighted: none of them was a poem, short story, or any other genuine intellectual work. Add to this that all comments are replies to a post, i.e., they do not constitute a full story by themselves.

Summary

Chapter 3 introduced the research philosophy, highlighted main research methodologies, and how they were used in this research. The chapter then explained the data collection process and the ethical frame governing it. The chapter ends by addressing different ethical consideration, zooming out to explain IP and copyright in general, and providing guideline to follow during before data collection and usage.

The chapter highlighted that in addition to ensuring the legal and ethical aspects of data collection and usage were considered, the emotional and physical wellbeing of users whose comments were used was a priority. None of the stages of the work had evaded the privacy of users or revealed information about them that may cause them any harm. No contact, personal, or geographical information of the users were collected or used neither were their profiles visited in the first place. Only the comments that were posted under public mode and as replies to the pages' owners were used. Finally, no commercial or morale benefits were achieved out of the comments, neither any comment may be copyrighted by the users later on. All the comments are informal phrases written in IA and none of them constitutes a work of art.

Moreover, a clear ethical framework ensured that no copyrights violations took place at any stage of this research, and ensured the physical, emotional, and financial wellbeing of all participants and researchers involved. In general, an important starting point prior to data collection and usage is following the academic institution's guidelines and consulting the ethics committee. Afterwards, one needs to check the terms and conditions of any institution or organization involved besides ensuring the all guidelines related to copyrights are followed, a good source to such guidelines can be found at WIPO's website. Appendix C has some useful info related to FB's terms and conditions and the ethical committee checklist.

CHAPTER 4: Development of Sentiment Resources and Sentiment Classifier

The core aim of this research is to classify sentiment expressed in IA and develop insights and understanding from it. Chapter 4 tackles two of the research objectives that are considered two building blocks necessary to address this core aim: constructing the annotated corpus and constructing the sentiment lexicon. It also prepares the stage for implementing the classifier by specifying the rules to be followed in classification; the classifier's design will appear in chapter 5.

4.1 Introduction to Classification using Lexicon

Each language (and dialect) has specific words or statements that are used to express positive and negative opinions. Specific statements are said in case of condolences, weddings, sarcasm, cursing, congratulations, etc. Based on this, this research aims to study how online users use these words and statements, which will be called lexemes hereafter, to express a sentiment. A lexeme is either a standalone word that is enough to express a sentiment such as قبيح (which means ugly), or a phrase that has a sentiment when used as a whole without having any opinionated word such as القرد في عين امه غزال (which is a proverb that means a monkey always seems beautiful to its mother). Afterwards, lexemes were normalized in a way to increase recall as much as possible. For example, the lexeme منافقة (which means hypocrite) is used to describe a female. However by removing the last letter, the lexeme will be used to describe a male.

Facebook comments were collected and studied to see which lexemes are usually used to express opinions and how these lexemes are different than those used in MSA. In the next section, we describe the corpora constructed for this purpose and how lexemes were used to classify comments. The constructed corpora consist of comments written in IA, the current classification is done per sentiment and not per dialect. In other words, the classification dealt with all dialects without differentiating them, first because the original objective is to know the sentiment of text and not its dialect, and second because no correlation between sentiment and dialect was found in our corpus. Finally, the implemented approach cannot be considered a regular bag of word approach because the frequencies of terms in a comment are not used as features, only the

occurrence of positive, negative, or spam lexemes, regardless of the frequency was used to classify the posts. Another difference from the typical bag of words approach is the handling of negation: the order of words does make a difference in our approach, as the presence of inverters flips the polarity of the opinionated lexemes following them.

4.2 Building the Corpus

Social media users post textual data that contain their opinions towards different objects (policies, institutions, products, etc.) ((Itani et al., 2017a, 2017b). The huge size and noisy nature of these data make manual extraction and classification of these opinions an infeasible task. For this purpose, classifiers are needed that can automatically extract and classify these opinions. The literature mentions five different classes that a sentence (or document) may have: negative (expressing negative sentiment like aggressiveness or sadness), positive (expressing positive sentiment like optimism or happiness), dual (also called mixed, containing both negative and positive sentiment), spam (advertising for an object) or neutral (which is informative text with no sentiment). There are various kinds of classifiers such as NB, DT, SVM, kNN, etc. Each of these classifiers follow a specific algorithm to classify text (sentence or document) as one of the classes mentioned earlier. However, to test these classifiers, and in supervised learning context, the classes of the text used to train/test the classifier should be known. It is usually specified by native speakers of the language who read the text and classify it according to predetermined set of rules. Before demonstrating how the corpora provided by this research were constructed, first, we summarize some of the currently available corpora. Then we describe the data collection process. Afterwards, the pre-processing applied on collected data is discussed. Finally, we describe different actions done during manual classification.

4.2.1 Current Corpora

Different NLP applications such as text categorization, machine translation and SA require a corpus to be used by the applications being developed. For example, if a set of opinionated online documents is to be classified, then annotated documents should be available to train and test the implemented classifier. Documents such as movie or product reviews available online are good examples of such opinionated documents. In context of SA, documents are classified according to their sentiment (positive, negative, neutral, etc.). Hence, classifiers are needed that can automatically extract and classify these opinions (Abbasi et al., 2008; Maynard & Funk,

2011; Farra et al., 2010; Hamouda & Akaichi, 2013; Itani et al., 2017a). To train and test such classifiers, corpora are needed. To test our LB classifier, we prepared our own corpora from Facebook comments written in IA. Table 3 briefly mentions some details related to different samples corpora used in literature, mainly focusing on their source, size, language, availability, and level of annotation (column SA Level) with S standing for sentence, W for word, and D for document. For instance, if the corpus consisted of annotated documents, the value of SA Level will be D. For the sake of proper readability, the sources to corpora in table 3 are shown in table 4. It is worth mentioning corpora below are not directly related to our work and not all of them can be used for SA, but are mentioned to show a sample of the corpora that exist in literature in terms of source, size, language, annotation level, and domain.

Table 3 – Examples of Different Corpora and their Properties

ID	Source	Size	Language	labelled	Classes	Availability	Domain	SA level
1	Facebook	2000	IA	Yes	5	Yes	Arts/News	D
2	Yahoo!	1100	English	Yes	N/A	Yes	misc.	S
3	Twitter	8868	MSA/CA/IA	Yes	4	yes	misc.	S
4	PubMed	9985	English	Yes	3	No	Medicine	S
5	Newspapers	104	Brazilian/Portuguese	Yes	N/A	Yes	news	D
6	Essays	120	Japanese	Yes	N/A	No	Arts	D
7	Recordings	1.5M Words	Dutch	Yes	N/A	Yes	misc.	W
8	Journal	385	English	Yes	N/A	Yes (Fee)	misc.	D
9	Webpages	1000	Japanese	Yes	N/A	No	misc.	D
10	Researches	1434	English	Yes	N/A	No	Engineering	D
11	Newspapers	102134	MSA	No	N/A	No	News	D
12	Webpages	2232	Arabic/English	No	No	No	Science	D
13	misc.	6M Words	MSA	Yes	N/A	yes	misc.	W
14	UN Documents	3M Words	Arabic/English/Spanish	Yes	N/A	Partly	misc.	W
15	Written Docs	0.25M Words	Swedish/Turkish	Yes	N/A	No	fiction/technical	W
16	AQUAINT-2	2.4GB	English	Yes	N/A	No	Motion-Specific	D
17	ATB2 v 3.1	501	MSA	Yes	N/A	No	news	D
18	N/A	36,895	IA/English	Yes	N/A	No	misc.	S
19	Twitter	52000	Arabic/French	No	N/A	No	news	S
20	Webpages	28,530	Italian	Yes	N/A	No	misc.	D
21	Webpages	9.7M Words	Arabic/English/Swedish	No	N/A	No	IT	W
22	Webpages	400000	Chinese/English	Yes	N/A	No	misc.	S
23	Webpages	N/A	English	Yes	N/a	No	N/A	D
24	Webpages	80000	11 Euro Languages	No	N/A	yes	Politics	D
25	Twitter	50,324	N/A	No	N/A	yes	misc.	S
26	Webpages	22429	MSA	No	N/A	yes	misc.	D
27	Webpages	1M Words	MSA	No	N/A	yes	misc.	D
28	Social Networks	15372 Words	MSA	No	N/A	No	N/A	W
29	misc.	6M Words	MSA	Yes	No	No	misc.	W
30	ATB1V3	400	MSA	Yes	4	No	misc.	D
31	Quran	77430 Words	CA	Yes	N/A	Yes	religious	W
32	Written Docs	16329	MSA	Yes	N/A	yes	questionnaire	W
33	Webpages	20291	MSA	Yes	N/A	yes	misc.	D
34	Webpages	500	MSA	Yes	2	yes	Movie Reviews	D
35	Webpages	1M Words	MSA	Yes	2	yes	News	W
36	misc.	N/A	MSA	No	N/A	No	misc.	D
37	Social Networks	14993	MSA	Yes	4	No	misc.	S
38	ATB1V3	2855	MSA	Yes	4	yes	misc.	S

39	Webpages	44	MSA	Yes	3	No	Movie Reviews	S/D
40	Facebook	260	English	Yes	2	No	Politics	S
41	Facebook	6000	MSA	Yes	3	No	misc.	D
42	Webpages	1143	MSA	Yes	2	No	misc.	D
43	Webpages	2000	English	Yes	2	Yes	Movie Reviews	D
44	misc.	37M Words	MSA	No	N/a	No	misc.	W

Table 4 - Source of Corpora mentioned in table 3

ID	Corpus
1	(Itani, 2017)
2	(Atserias et al. 2010)
3	(Refaee and Rieser, 2014)
4	(Houngbo and Mercer, 2014)
5	(Caseli et al., 2009)
6	(Iida and Tokunaga, 2014)
7	(Oostdijk, 1999)
8	(Carlson et al., 2003)
9	(Hangyo et al., 2012)
10	(Liu et al., 2004)
11	(Abdelali et al., 2005)
12	(Mustafa and Suleman, 2011)
13	(AbdelRaouf et al., 2010)
14	(Samy et al., 2006)
15	(Megyesi et al., 2006)
16	(Roberts, 2009)
17	(Bahloul et al., 2014)
18	(Riesa, et al., 2006)
19	(Hajjem et al., 2013)
20	(Baroni and Ueyama, 2006)
21	(Izwaini, 2003)
22	(Baobao, 2004)
23	(Aleahmad et al., 2009)
24	(Koehn, 2005)
25	(McCreadie et al., 2012)
26	(Saad and Ashour, 2010)
27	(Al-Sulaiti and Atwell, 2006)
28	(Akra, 2015)
29	(Al-Sabbagh and Girju, 2012)
30	(Abdul-Mageed and Diab, 2012)
31	(Dukes and Habash, 2010)
32	(Rytting et al., 2014)
33	(El-Haj and Koulali, 2013)
34	(Rushdi-Saleh et al., 2011)
35	(Maamouri et al., 2004)
36	(Alansary et al., 2007)
37	(Abdul-Mageed et al., 2014)
38	(Abdul-Mageed et al., 2011)
39	(Farra et al., 2010)

40	(Hamouda and Akaichi, 2013)
41	(Hamouda and El-Taher, 2013)
42	(El-Halees, 2011)
43	(Pang et al., 2002)
44	(Zemánek, 2001)

Placing this work in context, it is distinctive in terms of source used, extraction method, size, and classification.

4.2.1.1 Sources of Corpora

Upon checking different sources used to build corpora, we note that the literature can be roughly divided into seven categories as follows:

1. Corpora that were built by mining data from databases: (Iida and Tokunaga, 2014; Hounbo and Mercer, 2014; Carlson et al., 2003; Liu et al., 2004; Samy et al., 2006; Dukes and Habash, 2010)
2. Corpora that were manually built by treating written text: (Megyesi et al., 2006; Rytting et al., 2014)
3. Corpora built by downloading and processing webpages: (Caseli et al., 2009; Hangyo et al., 2012; Abdelali et al., 2005; Mustafa and Suleman, 2011; Baroni and Ueyama, 2006; Izwaini, 2003; Baobao, 2004; Aleahmad et al., 2009; Koehn, 2005; Saad and Ashour, 2010; Rushdi-Saleh et al., 2011; Maamourir al, 2004; El-Halees, 2011; Pang, et al., 2002).
4. Corpora built by downloading social network posts: (Atserias et al., 2010; Refaee and Rieser, 2014; McCreadie et al., 2012; Akra, 2015; Abdul-Mageed et al., 2014; Hamouda and Akaichi, 2013; Hamouda and El-Taher, 2013).
5. Corpora built by downloading subset of a Treebank: (Roberts, 2009; Bahloul et al., 2014; Abdul-Mageed and Diab, 2012; Abdul-Mageed et al., 2011).
6. Corpora build from multiple sources (different combination of the upper sources): (AbdelRaouf et al., 2010; Al-Sabbagh and Girju, 2012; Alansary et al., 2007; Zemánek, 2001).
7. Corpora built by recording voices: (Oostdijk, 1999)

Since our work focuses on textual data, we will focus on the first six categories. This work is similar to those mentioned in the fourth category. Specifically, the closest data type to ours is that described in (Hamouda and Akaichi, 2013; Hamouda and El-Taher, 2013) since the same social network, Facebook, is used as a data source.

4.2.1.2 Data Extraction

To download data from online sources, three ways can be used:

1. Using Application Programming Interface (API): Social media such as Twitter provide an interface that allows automatic download of posts, this was used in many researches (Refaee & Rieser, 2014; Hajjem et al., 2013; McCreadie et al., 2012). Such applications are specific to one social network and do not exist for all social networks.
2. Crawling: crawlers can be used to automatically download data of target webpages, using this procedure will require further processing to remove noise such as timestamps, html tags, etc., (Abdelali et al., 2005; Roberts, 2009; Baroni & Ueyama, 2006; Izwaini, 2003; Baobao, 2004; Koehn, 2005; Saad & Ashour, 2010; Al-Sulaiti & Atwell, 2006; Akra, 2015; El-Haj & Koulali, 2013; Maamouri et al., 2004; Abdul-Mageed et al., 2014; Itani et al., 2017). Crawling is usually used to build corpora of large sizes where manual extraction is infeasible.
3. Manual Download: This kind of data extraction suits small to medium size corpora. Online data can be saved and processed to fit the classifier being used. This technique was used in many researches (Abdelali et al., 2005; Roberts, 2009; Baroni & Ueyama, 2006; Izwaini, 2003; Baobao, 2004; Koehn, 2005; Saad & Ashour, 2010; Al-Sulaiti & Atwell, 2006; Akra, 2015; El-Haj & Koulali, 2013; Maamouri et al., 2004; Abdul-Mageed et al., 2014; Itani et al., 2017a, 2017b) and is the one adopted by us to build our corpus. It is worth mentioning that data collection respected Facebook's terms and conditions and no copyright infringement took place in the process.

4.2.1.3 Corpora Size

Corpora size reported in literature used either number of words, number of sentences, or number of documents. We consider each record in our corpus as a document since each post is of arbitrary length with no specific limit. Therefore, we compare our work to others who also used a number of documents to measure their corpus size.

Corpora Size Reported in Terms of Number of Documents:

Small Corpora ranged from 44 to 6000 documents while large corpora ranged from 20291 to 102134 documents.

4.2.1.4 Corpora Annotation

In addition to prepare the raw corpus, additional metadata can be added to the corpus depending on how the corpus is to be used, based on our review, we were able to categorize the corpora as being annotated or not, with our corpus being annotated:

Annotated Corpora: (Atserias et al., 2010; Refaee & Rieser, 2014; Houngho & Mercer, 2014; Caseli et al., 2009; Iida & Tokunaga, 2014; Oostdijk, 1999; Carlson, 2003; Hangyo et al., 2012; Liu, Loh, & Tor, 2004; AbdelRaouf, et al. 2010; Samy et al. 2006; Megyesi et al., 2006; Roberts, 2009; Bahloul et al., 2014; Riesa et al., 2006; Baroni and Ueyama, 2006; Baobao, 2004; Aleahmad et al., 2009; Dukes & Habash, 2010; Rytting et al., 2014; El-Haj & Koulali, 2013; Rushdi-Saleh et al., 2011).

Unannotated Corpora: (Abdelali, Cowie, & Soliman, 2005; Mustafa & Suleman, 2011; Hajjem et al., 2013; Baroni & Ueyama, 2006; Koehn, 2005; McCreadie et al., 2012; Saad & Ashour, 2010; Al-Sulaiti & Atwell, 2006; Akra, 2015; Alansary et al., 2007; Zemánek, 2001).

As for the number of classes adopted in each corpus, the number and type differ from one corpus to another. Of those examined, the number of classes used, number of classes ranged from 2 to 5 pre-determined categories. Other kinds of classification do exist (Carlson et al., 2003) where authors used 16 different classes of annotations, however, such annotations are mainly used as metadata describing each textual unit (word, sentence, document) and their characteristics.

Generally, annotation is done by native speakers based on predetermined rules. The manual annotations are later used to test the accuracy of the classifier. As for the quality of these annotations when more than one annotator did the tagging, it is decided based on majority vote or by randomly choosing one of the disagreed-on labels when the votes are even.

In our case, we only adopted annotations where all annotators gave the same annotation. Section 4.2.4 will go through our annotation process in details. The corpus is available for public use (Itani, 2017).

4.2.2 Data Collection

El-Haj et al., (2015) described three approaches to build resources for under-resourced languages: (1) using crowdsourcing, (2) translating an existing corpus, and (3) constructing a corpus manually. In this work, their third approach is adopted, i.e. using manual effort with skilled experts to collect and annotate the corpus. The constructed corpora were built using Facebook textual comments. The size of a comment ranged from one word to a document containing many sentences. We chose Facebook since it is the social network with the biggest number of users, 1.11 billion users (“Facebook Company Statistics”, 2017), it is the one preferred by Arabs (“Facebook in the Arab Region”, 2017), and because it allows comments of larger sizes than other social networks such as Twitter whose post size is formally (“Twitter Developer Documentation”, 2017).

Two corpora were built of two different domains, arts and news, to be used later on in SA or in domain classification. 1000 comments were collected from Al Arabiya News FB page (will be called NC hereafter) (Al Arabiya, 2012) and 1000 comments were collected from The Voice Facebook page (will be called AC hereafter). The Voice is a singing competition (MBCTheVoice, 2012). The comments consist of textual data posted by users in response to posts written by the pages’ owners. A sample of comments can be found in appendix B. The majority of these comments (~95%) were written in IA and not in MSA, this was assessed during manual classification of the comments. This confirmed our expectation regarding the use of MSA in social media.

To help reduce the risk of a ML classifier being over influenced by high frequency classes, the annotators kept track of the frequency of each class to ensure the corporate was roughly balanced across classes.

The reason behind choosing two domains, arts and news, was to check if the same classification technique worked effectively in different domains. The similar classification results for each suggest that sentiment classification is not strongly domain biased. However, classification does benefit from domain-specific knowledge as discussed by Alfrjani et al. (2016) and Aljamel et al. (2015)

4.2.3 Pre-processing

After data collection, comments were pre-processed on three different stages:

- (1) Removing redundancies: online users tend to post the same text more than one time in the same thread, either to show passion towards an object (like cheering for an artist), to show objection towards a topic (using curse words and offensive language), or to spread a spam, i.e., to post a hyperlink referring to another website or Facebook page and inviting users to visit that page.
- (2) Removing time stamps: each comment has a timestamp that mentions when the comment was written, in context of SA, timestamps are of no significance, and therefore we removed them from collected comments.
- (3) Removing Likes: A like in Facebook terminology is a link that can be clicked by users to show that users like what has been posted. A group of emoticons may also be associated with a comment such as angry face or sad face, this work does not address the effect of the likes and emoticons on sentiment classification. A sample of the comments collected is shown in figure 4.1:



Figure 4.1 – Sample of Downloaded Comments

4.2.4 Manual Classification

Following data processing, four expert native Arabic speakers classified the collected comments. Each human annotator could read, write, and speak MSA in addition to having a good understanding of other Arabic dialects. All annotators are Lebanese, had a master's degree in different domains, educators at different levels and institutions, and one has a PhD in Arabic linguistics. All annotators are familiar with Egyptian, Syrian, and Palestinian dialects, and one

was also familiar with Iraqi. Any comment with vague meaning was considered to be unfamiliar, the experts consulted native speakers of the relevant dialect, such as: Egyptian, Iraqi, and Tunisian. The comments annotators are the author's friends and ex-colleagues, and are different than the 100 Facebook friends that contributed to the lexicon as mentioned in section 3.3.

In order to strengthen the validity of the manual annotation, only comments on which all four annotators agreed were added to the corpora, others were discarded. The process continued until the target of 2000 annotated comments with IAA of 100% was achieved. It is worth mentioning that the human annotators did not depend on the original post to classify the comments, they only classified the comments as if they were standalone posts and not comments on a posts. Therefore, the classification of comments was not affected by the sentiment of original post or its content.

Manual classification followed the rules below:

Negative: if the comment expresses negative sentiment or feeling such as sadness, pessimism, hostility or any other negative feeling. For example, *للأسف كان ذلك على حساب يسرى* (unfortunately, that was on Yusra's expense)

Positive: if the comment expresses positive sentiment or feeling such as enthusiasm, happiness, optimizing, etc. For example, *مراد مبروك* (congratulations Murad)

Dual: if the comment expresses negative and positive sentiments regardless of the frequency of each. For example, *مراد أخذ اللقب عن جدارة واستحقاق وموتوا بغيطكن يا حساد* (Murad deserves the title, die haters)

Spam: if the comment is inviting users to join or “Like” a Facebook page or to advertise. For example, *السلام عليكم ممكن تنشرون هذا البيج* (greetings, can you spread this page)

Neutral: if the comment is informative or expressing no sentiment. For example, *مراد شو شعورك ان ربحت احلى صوت وشو شعورك ان خسرت؟* (Murad how would you feel if you win or lose the competition?)

4.2.5 Corpora Characteristics

We tried to make the comments in each corpus balanced by collecting the same number of comments of each class. This was achieved during the collection and annotation phases by

replacing the comments whose class has higher frequency than other does, by comments whose classes has lower frequency. For instance, once the annotators found they had too many negative comments, they started disregarding new negative comments and replacing them with comments of lower class frequency. Table 5 shows the number of comments of each class in each corpus:

Table 5 - Frequency of Comments of Each Class

	AC	NC	Total
Negative	224	230	454
Positive	233	236	469
Dual	151	161	312
Spam	197	193	390
Neutral	195	180	375
Total	1000	1000	2000

AC contains 12053 words with an average of 12 words per comment whereas NC contains 8423 words with an average of eight words per comment.

4.3 Sentiment Lexicon

A lexicon resembles a dictionary in the sense that each entry is assigned a label, which is not necessarily the meaning of the entry. In a dictionary, each word is assigned a meaning. However, in a lexicon, each entry, also known as lexeme, may be assigned a label (negative, positive, domain (science, sports), or any other chosen label depending on what the lexicon will be used for.

4.3.1 Importance of Sentiment Lexicon

SA uses semantic, stylistic and syntactical features. Semantic features include opinionated lexicons. Such lexicons usually belong to two different classes, negative and positive. LB classifiers classify text depending on the presence of sentimental lexemes (such as good and bad). The presence of positive lexemes indicates positive sentiment, presence of negative lexemes indicates negative sentiment, and presence of both indicates mixed or dual sentiment. We try in this work to provide additional resource for SA of IA by providing three sets of lexemes, negative, positive, and spam.

4.3.2 Building the lexicon

In section 2.2, we highlighted works related to SA, now we zoom in to focus on works related specifically to building lexicons. There are numerous works in this area and they differ over

source, size, language of annotation, etc. Building lexicons falls roughly into three main categories (Banea et al., 2008; Badaro et al., 2014; Lu et al., 2013; Lu et al., 2011; El-Abbadi et al., 2013; Tsunakawa et al., 2008; Dzikovska, et al., 2004; Bamman et al., 2008; Tang et al., 2014; Olteanu et al., 2014; Wilson et al., 2005; Abdul-Mageed & Diab, 2014; El-Beltagy, 2016):

- A. Manually compiled/annotated lexicon such as Harvard Inquirer, WordNet(s), Micro-WNOp.
- B. WordNet-based approaches, with and without scores such as SentiWordNet.
- C. Multi-source Lexicon where authors collect their lexicons from different sources such as dictionaries, manual labelling, and online sources.

Our research followed approach C, i.e., different sources were used to construct the lexicon. The sources are (1) manual extraction, (2) surveying, and (3) extracting words from the dictionary.

4.3.2.1 Manual Extraction of Lexemes

In addition to giving a class for each comment, the human annotator extracted lexemes from each comment, lexemes that according to the human annotator were behind the giving the comment its class. For example, مراد مبروك (congratulations Murad)

The word مبروك (which means congratulations) was extracted and added to set of positive lexemes. Although the comment could be the full name of a person, annotators knew from context that it is not because “Murad” was the first name of a competitor in The Voice and “Mabrook” was the Arabic equivalent of “congratulations”. This reflects what is mentioned by Alfrjani et al. (2016) concerning effect of domain knowledge on automatic classification performance. It is worth exploring such issues further, and whether other NLP tools can improve upon such challenges. In this case, Named Entity Recognition (NER) was conducted using MADAMIRA. MADAMIRA is a system that has different Arabic NLP tools such as NER, POS tagging, and tokenizing (Pasha et al., 2014). MADMIRA is currently suitable for MSA and Egyptian dialect it tagged both words as nominal (when Egyptian dialect was used) and not a proper noun, which is in keeping with the annotators tag. It is worth mentioning that MADAMIRA’s NER gave the correct tag when MSA was used and not when Egyptian Dialect. In the context of examining social media, we cannot in general assume comments to be in MSA. The NER output is show in figure 4.2:

Standard Arabic

▶

Named Entities

مراد مبروك

Parts-of-Speech

Tokenized Forms

Diacritized Forms

Lemmas

Base Phrases

Named Entities

مبروك

مراد
Person

Figure 4.2 – Sample Output from MADAMIRA's NER

Another two examples where the NER and POS tagger failed to determine the proper noun correctly are found below (figures 4.3 and 4.4). Such cases emphasise that NLP tools can significantly contribute to SA, and that their performance is still not accurate enough when IA is used.

In the first, the POS tagger tagged قصي (Qusay) as a verb, where in fact it is a proper noun.

قصي يا رائع

Parts-of-Speech

Tokenized Forms

Diacritized Forms

Lemmas

Base Phrases

Named Entities

قصي يا رائع

verb nominal particle proper noun

Figure 4.3 – POS Tagging Sample Output 1

A similar issue appeared when the POS tagger tagged مبروك (incorrect spelling of the Arabic equivalence to congratulations), as a proper noun where in fact it is an adjective.

مبروك النجاح

Parts-of-Speech

Tokenized Forms

Diacritized Forms

Lemmas

Base Phrases

Named Entities

مبروك النجاح

verb nominal particle proper noun

MADAMIRA in Arabic

MADAMIRA in English

باللغة العربية

باللغة الإنجليزية

References:

مبروك

POS: Proper Noun

Gender: Masculine

Number: Singular

State: Indefinite

Gloss: NO ANALYSIS

Figure 4.4 – POS Tagging Sample Output 1 2

It is worth mentioning that MADAMIRA’s POS tagger and NER operate under two modes, MSA and Egyptian Dialectal Arabic, and both failed to tag the posts mentioned above.

The set of lexemes extracted from the news comments was called NL and the set of lexemes extracted from arts comments was called AL. Lexemes were extracted from the comments regardless of the polarity of the comments, in other words, even if the comment was spam for instance and at the same time containing positive (or negative) lexemes, these lexemes were extracted and added to the corresponding set. By way of example, انت احلى صوووووووووووت (You have the most beautiful voice) is positive because it contains the positive lexeme “beautiful”; hence this lexeme was added to the set of positive lexemes.

After extracting lexemes, normalization was applied to guarantee higher recall. For the time being, we will define recall as the ability of the classifier to detect presence of a sentiment lexeme and hence increase its classification performance, detailed definition will appear in chapter 6. Normalisation was done in a two-phase process:

- a) Removing repeated letters.
- b) Convert extracted lexemes to regular expressions (“Regular Expressions Info”, 2017): Lexemes were extracted during manual classification of comments and were added to the corresponding set of lexemes (negative, positive, spam). It was noted during manual classification that many lexemes are different variants of the same word, for example, the word مبروك (which means congratulations) is written as ممممبروك or مبروووووووك, these two words are two incorrect variants that contain repeated letters. It is worth mentioning that this type of spelling variation falls under intensification. Online users tend to repeat random letters in sentimental words to boost or amplify their sentiment. In English for instance, an online user would write “goood” or “goodddd” instead of “good”. A regular expression can detect if the original lexeme is present in a spelling variant as long as the order of letters is not altered; to use the example of the lexeme “good”, as long as g’s come before at least two o’s, and then followed by d’s, then the regular expression can detect that “good” was present. Normalization excluded cases that may lead to ambiguity such as بكرها (which means I hate her), since removing the last letter will result in بكره (which may mean “tomorrow” or “I hate”) which may mean “I hate him” or “tomorrow”. In this context, normalising means using a lexeme that can replace many other lexemes and therefore reduce search time. For instance, the lexeme “play” can replace all different variants such as “played”, “plays”, and “playing”. So instead of searching for four entries “play”, “plays”, “played”, and “playing”,

only one is searched for, the one that is present in all variants, in this case “play”. The conversion from lexemes to regular expression has been automated in our system so any use of a new lexicon needs to consider the lexemes carefully, keeping in mind that all the regular expressions will be ignoring letter repetitions and not doing morphological analysis, i.e., in case a verb or an adjective changes significantly according to tense its tense, then both version of the lexemes need to be fed. For instance, “break” and “broken” are considered two different lexemes, unlike “break” and “breaks” because in the former case the two words are totally different whereas in the second, “break” is totally contained in “breaks.”

It is important to mention that this work uses one feature of regular expression which is its ability to detect repetition of letters, and that repetition here refers to those that are not original part of the word, for instance, consider the word ممنوع (Forbidden), that has a letter repeated twice, the regular expression in this case will check any variants where this letter is repeated more than twice, and considers them as being the same lexeme. In effect, if this use of regular expressions to normalise lexemes is effective, it captures one element of the informality of IA. Repeated letters correctly spelt or otherwise are irrelevant to sentiment.]

In addition to applying the manual extraction on the arts and news corpora, the same process was applied on two other corpora with the aim of extracting sentimental lexemes from them; the two corpora are Anew and Nnew:

Anew represents a corpus of arts comments different from those of AC.

Nnew represents a corpus of news comments different from those of NC

AL2 represents the number of lexemes extracted from Anew.

NL2 represents the number of lexemes extracted from Nnew.

Characteristics of Anew and Nnew are of no significant as they were only used to extract new lexemes.

4.3.2.2 Surveying

To strengthen the range of negative and positive phrases an open request for sentimental lexemes was posted on FB. The users were informed that their comments will be used for research purposes and the two threads used were deleted after data collection. The post used to collect

negative phrases comments could be reasonably translated as: “For research purposes, please comment with a word or phrase that you would use to curse or to express negative feeling such as anger and sadness.” A similar post was used to collect the positive lexemes. Afterwards the comments were manually checked and redundancies were removed. Overall, 538 lexemes resulted from the activity.

4.3.2.3 Extracting Lexemes from the Dictionary

Arabic dictionary was used to boost the lexicon by adding lexemes that express negative or positive sentiment. This extraction is partial since the Arabic language has hundreds of thousands of words, only a small portion was chosen. Revisiting all words of the dictionary is part of our future work. Extracting the lexemes from the dictionary did not follow a specific algorithm neither a saturation point was pre-set, the dictionary was randomly searched for opinionated lexemes.

The output of lexicon construction described is shown in table 6. In the upcoming section, the whole lexicon name total in table 6 will be referred to as Gold.

Table 6 - Numbers of Lexemes in the Lexicon Grouped Per Source

	Spam	Positive	Negative	Total
AL	73	516	705	1294
NL	42	531	666	1239
AL2	92	942	939	1973
NL2	11	150	698	859
Surveying	0	223	315	538
Dictionary	0	2365	1559	3924
Total	218	4727	4882	9827

Moreover, it was noted in our corpora that spam lexemes are dominant because they always override other lexemes in a comments, i.e., if a comment has a spam lexeme and a positive (or negative) lexeme, the comment was found to be a spam according to the manual classification. This dominance property of spam lexemes affected the way the classifier was implemented by giving spam lexeme a priority in classifying comments as will be shown in next chapter.

Table 7 shows the percentage of each dialect of the total number of lexemes. The category “Common” refers to cases where it was not possible to determine the dialect if the phrasing of the comment is common to many dialects.

Table 7 - Percentage of each Dialect in the Lexicon

Dialect	Percentage
Levantine	1.22%
Egyptian	2.98%
Iraqi	0.63%
North Africa	0.46%
Common	94.71%

We now analyse variation of lexemes polarity per dialect (Table 8). Although dialectal lexemes were not frequent, we noticed in our lexicon that within the same dialect, the majority of lexemes were positive. However, the dialectal lexemes in our lexicon are few (5.29%), and therefore it is not possible to generalize. One possible reason for the dialectal lexemes to have more positive entries than negative ones would be that negative lexemes including cursing words are common to all dialects, or there may be other social, cultural, or psychological reasons that make online users use common negative lexemes throughout their online conversations, this issue remains open and needs further analysis from NLP, social, and psychological perspectives.

Table 8 - Percentage of Lexemes in Dialects

Dialect	Spam	Pos	Neg
Levantine	17.24%	75.86%	6.90%
Egyptian	5.63%	76.06%	18.31%
Iraqi	0%	80.00%	20.00%
North Africa	0%	81.82%	18.18%
Common	4.03%	52.13%	43.84%

4.4 Negation

Since the constructed lexicon contains IA text, we couldn't assume that the inverters used are those used in MSA (لا, ما, لم...) and we had to treat each case separately. We went through the extracted lexemes and filtered all those containing inverters. Afterwards, we checked the polarity of targets and added them to their corresponding sets. For example, if a negated phrase had a positive target, this target was added to the set of positive lexemes. However, there are plenty of cases in which negation exists; yet the target alone is meaningless and does not express a sentiment. For example, لا صوت ولا صورة (exact translation is: no sound and no picture) is a negated phrase that is inverting two neutral targets. Yet, when negation is applied on these two neutral nouns, the meaning becomes negative and indicates that someone is ugly and cannot sing. Such lexemes were not split and were kept as they are. Afterwards, the algorithm of

classification was modified to consider negation: for the first positive (and negative) lexeme found, the previous word was checked to see if it is an inverter, and then if an inverter was found the polarity of the lexeme was inverted. Surprisingly, the results after treating negation was against expectation. The expectation was to have a significant improvement in performance, but this did not happen due to the complexity of negation in IA for several reasons. One example of such reasons is concatenating the inverter to the target: for example the lexeme محب, which means “a lover” in MSA, but may mean a “lover”, or “he did not like” in informal Arabic. Additional reasons such odd negation and fake inverters will be discussed in details in chapter 6.

Specific words are used in Arabic to negate targets. This negation may flip the polarity of opinionated words as in ليس جميل (not beautiful). In MSA, these words are limited (لا, لم, لن, etc.) and can be easily detected because when spelling rules are properly used. However, in IA, these words change according to the dialect, and since no spelling rules can be enforced, detecting such words is a harder task. Table 9 lists inverters used in MSA and table 10 lists some of the inverters used in IA, the ones found in our corpus. MSA inverters can be used in IA but the opposite is not true. The last IA inverter, م, acts as a suffix and negates the target to which it is attached, such as محببتو (I did not love him). The second and fourth columns are close translations of the meanings of the MSA and IA inverters to English, however, the meaning could differ according to context.

Table 9 – Common MSA Inverters

Inverter	Meaning
لا	no
لم	did not
لما	did not
لات	no
لن	will not
بلا	without
ليس	not
من دون	without
بدون	without

Table 10 - Common IA Inverters

Inverter	Meaning
مفي	there isn't any
مافي	there isn't any
منو	not
ماكو	there isn't any
مو	not
بلاش	without
مش	not
مانو	not
*م	not

The inverters can appear as a separate word directly preceding targets as in جميل (not beautiful) or part of the negated word as in محببتي (I did not like him). Inverters do not necessarily flip a positive (negative) sentiment of a target into a negative (positive) sentiment. Consider the following negated phrase: ما تهزأ (don't make fun of). The verb has a negative sentiment, yet it is still negative after being preceded by an inverter. The same inverter can be used to flip polarity of opinionated words such as (ما تزعل) (don't be sad). Although the same inverter is acting on verbs in both cases, it has two different effects. The same string ما can be used to express meanings not related to negation: it can be used to ask questions as in ما اسمك (what is your name), or to praise a target ما احلى صوتو (what a beautiful voice). Given the ability of inverters to flip polarity if sentiment, we modified the classification algorithm to cope with this effect: for each positive or negative lexeme, the lexeme's polarity was flipped if preceded by an inverter.

Summary

The chapter described the classification process and the construction of two of its building blocks: the corpus and the lexicon. The chapter introduces LB and describes available corpora and compares them against the corpus developed as part of this research. Afterwards, data collection, pre-processing, and corpus manual classification is described. Next, the chapter describes how the lexicon was constructed. It provides a detailed description about the lexemes used in classification and explains how regular expressions were used to increase classification recall when new corpora are to be classified. The major contribution of this chapter is to provide additional resources for SA of IA: annotated corpora and sentiment lexicon.

CHAPTER 5: Classifier Design

5.1 Introduction

Chapter 4 described how the corpora and the lexicon were built. An important feature of the lexicon was to include lexemes in the form of regular expressions. The corpus is needed to check the performance of the classifier by comparing automatic classification against manual one. The lexicon contains the lexemes that will be used to classify comments into one of five categories. This chapter puts all the pieces together and provides a high-level demonstration on how the classifier operates and performs.

The implementation is driven by various concerns:

- To allow for dynamic usage of the classifier for different domains.
- To allow for dynamic re-use with say different dialects.
- To support end users in the understanding and validation of classification results. For this purpose, a statistics summary was generated to show the distribution of different sentiments: percentage of objective (neutral) and subjective (positive, negative, dual, or spam) comments, the number of lexemes affected by inverters, and the most frequent lexemes.

The figures in appendix E illustrate how the domain's underpinning analyses are inputs that can be easily changed to suit the domain focus of any analysis.

This chapter satisfies the second research objective, which is to develop an LB classifier that uses a sentiment lexicon to classify SM comments written in IA according to their sentiment. It is considered achieved since it does provide one complete package that allows SA of IA. Section 5.2 describes the classification algorithm, section 5.3 illustrates how the corpus, lexicon, and inverters are uploaded, and section 5.4 illustrates the classification process.

5.2 Classification Algorithm

Since an LB classifier is being implemented, it was designed to mimic human classification of comments, i.e., to classify comments according to present of sentiment lexemes while considering that spam lexemes dominate negative and positive lexemes. However, as discussed in chapter 2, sentiment is affected by negation and the whole sentiment of a comment may

change due to the presence of inverters. As mentioned in section 2.7, this work only addresses the negation cases where inverters come directly before the words or phrases to be negated. The classification algorithm is shown below.

Declaration:

Numeric Spam-Counter = 0, Pos-Counter = 0, Neg-Counter = 0

String AutoClass = "Unclassified";

//ManualClass of a comment is the one given by the Manual
Tagger

Input: Unclassified Corpus C;

Output: Classified Corpus C;

Boolean: Spam-flag = false, Pos-flag = false, Neg-flag = false;

AutoClass in {Spam, Dual, Pos Neg, Neu}

For Every comment P

 If P has a substring that matches Spam Lexeme

 Spam-flag = true;

 Else

 For each Positive Lexeme in P

 If the Positive Lexeme is immediately preceded by
an inverter

 Neg-flag = true;

 Else

 Pos-flag = true;

 endif

 end for each

 For each Negative Lexeme in P

 If the Negative Lexeme is immediately preceded by
an inverter

 Pos-flag = true;

 Else

 Neg-flag = true;

 endif

 end for each

endif

Table 11 shows the classification truth table followed by the classifier.

Table 11 - Classification Truth Table

Spam-flag	Pos-flag	Neg-flag	AutoClass
T	-	-	Spam
F	T	T	Dual
F	T	F	Pos
F	F	T	Neg
F	F	F	Neu

5.3 Uploading Corpus, Lexicon, and Inverters

In total, the classifier needs five different files uploaded prior to classification: the comments (the corpus), the inverters, negative lexemes, positive lexemes, and spam lexemes (the last three files constitute the lexicon) (see figure E.1).

Once uploaded, the files need to be loaded. The loading feature allows accumulating content from different files of the same type (i.e. different files containing positive lexemes). It works as follows: once a file is uploaded, and before loading it (using its content), the user can choose whether to empty available data from previous or to add the content of the uploaded files to previously uploaded (see figure E.2). The system will automatically convert uploaded lexemes into regular expression prior to classification.

5.4 Classification Results and Statistics

After corpus and lexicon have been uploaded, the user can decide whether to classify the whole corpus or a specific subset of records. The user may also choose the number of records within a corpus to be classified (see figure E.3). Moreover, this can be done incrementally. The user can choose whether to add the classification results of specific records to results achieved in a previous classification of different records, or to clear previous results and start from scratch. We added the results accumulation functionality in case comparisons of results were needed as it supports the validation process.

When the “Start process” button is clicked, the classifier will search for matches between regular expressions of each type of lexemes (negative, positive, and spam) and the words in each record. When classification is done, a time stamp will appear indicating starting and ending times to support performance evaluation. Two layers of outcome analysis are provided: (i) filtering and grouping of classifications and (ii) statistical analysis to support the envisaged needs of end user

professionals as mentioned in the chapter introduction. For example, a marketing officer may be interested in seeing the negative comments only to steer the marketing campaign accordingly. Outcome analysis allowing filtering and grouping shows how classification took place for each record, including the number of lexemes found in each record, which lexemes were found, and whether inverters were used. This profiles each record. Although classification algorithm does not require to know the number of lexemes of each type but rather whether they were present or not, the information helps in assessing the process and investigating factor classifications such as dual posts. In addition, the number of lexemes affected by inverters is reported to help understand effect of negation on SA. Figure E.4 shows a sample output of the first reporting layer.

Users can focus on specific aspects of the analysis by grouping and filtering the classified comments according to criteria.

Grouping: allows user to group all records based on some criteria such as number of lexemes, class, presence of inverters, etc. For example, if we want to see all comments of similar polarity as one group (i.e. all negative comments in one group and all positive comments in another group, etc.). Moreover, grouping was implemented in a user-friendly manner; it is enough to drag the header of second column (Classification) into the grid header. Doing so will group all records into five groups (see figure E.5), clicking on any of the arrows will display records of the specified group only.

Filtering: each of the features used in the classification can be used as a filter, applying more than one filter at the same time is also possible. For example, to know which records included the pattern مبروك (congratulations), it is enough to write this pattern in the text box of positive lexemes as shown in figure E.6, and then only the records containing this lexeme will be displayed.

Finally, filtering and grouping can be done at the same time to cope with user's search preference.

Export: to download classification results, an export feature was added that can either download raw classification results, or results achieved after grouping or filtering the data. Exported file is in excel format to allow easy edit and view of data.

The second layer of reporting summarizes classification results. This may become useful whenever a large corpus is being classified. The following statistics were chosen:

- Frequency of occurrence of lexemes in each class (Positive, negative, spam, dual, and neutral).
- Percentage of lexemes affected by inverters.
- Percentage of each class.
- Frequency of occurrence of each lexeme.

It should be noted that the frequency does not represent the number of occurrences of the lexeme itself, but the number of times the regular expression representing the lexeme was able to match words of comments. For instance, if the counter indicates that the lexeme “congrats” has been found three times, then it means that either the exact matching of “congrats” or any of its spelling variants (with repeated letters) has been detected three times. Examples of the spelling variants include “congraaaaaats” and “congratttsss”. Figure E.7 lists for each class the number of lexemes that were detected upon classifying 101 comments. For example, in the comments that were classified as dual, 40 positive patterns were detected and 33 negative patterns (one of them was positive patterns but was flipped to negative because it was preceded by inverter). All Positive is the sum of positive lexemes (Original Positive), the negative lexemes that were negated (Flipped to Positive). All Negative is the sum of negative lexemes (Original Negative), the positive lexemes that were negated (Flipped to Negative).

Figure E.8 shows percentages and frequencies of each class, which can give a rough figure about the sentiment of the corpus. Figure E.9 shows the frequency of occurrence of each lexeme. The frequencies help decision makers in knowing what lexemes are contributing to the sentiments of the comments. It is worth mentioning that this report was added so that an analyst may check whether the frequent lexemes are in harmony with what the trend is within a community for expressing negative and/or positive opinions. For instance, if the word “cool” was among the popular words, yet it was not among list of frequent lexemes, this might indicate that the lexicon needs updating to include the word “cool.”

Summary

Chapter 5 discusses the classifier’s design, both conceptually and technically, covering the main functions and facilities. The classifier tool embodies the research and analysis of earlier chapters and enables the proposed classification approach to be tested and assessed for social media Arabic. The chapter also shows how the design can be driven by the way the classification results will be used. Different layers of reporting, filtering, and grouping may be added to cope with decision makers’ needs.

CHAPTER 6: Analysis and Validation

This chapter presents the primary results regarding the quality of the classifier described in chapter 5. The results are analysed in detail assessing them with respect to other existing NLP tools, and the original research questions. The chapter ends by focusing into analysing of spam comments and negation.

The detailed analysis is broken down as follows: section 6.1 provides primary results; sections 6.2 and 6.3 focus on same domain and cross-domain setups results; section 6.4 examines the effect of increasing the lexicon size; section 6.5 compares the results of NB and LB classifiers; section 6.6 analyses the classification results of using different lexicon on our corpora; and section 6.7 analyses classification results of our lexicon on different corpora.

Overall, the findings provide answers to the motivating research questions that demonstrate the relative value of the classifier developed as part of the research. Informally the LB classifier is shown to outperform a NB classifier if the NB classifier is only using n-grams without additional features.

The results also show that IA lexicon has the potential to give high results on diverse data sets. One of the lexicons used in the literature, NileULex ((El-Beltagy, 2016), outperformed our AL lexicon I one of the setups. Results also show that the increase in number of classes degrade classification performance.

Moreover, results show that the LB classifier should consider the number of classes existing in a corpus to avoid poor results. Using an LB classifier designed to classify data based on 5-point scale will not perform well if used on fewer number of classes.

Results also show the negation is complex in IA and that trivial solutions do not significantly increase classification performance. The complexity arises in terms of scope, homonyms, odd negation, and fake inverters.

Trying the lexicon on a corpus consisting of records from two different domains gave relatively good results, which shows that to some extent the lexicon is domain dependent, however, there is a room for improvement by using domain-specific knowledge.

6.1 Primary Results

To evaluate the efficiency of the classifier, the average F1-measure was used as defined in section 3.4. In addition, a new corpus CNew was also used to study how the LB classifier and lexicon will perform on new unseen corpus. CNew consists of 1000 unseen FB comments collected from the same two pages mentioned earlier. CNew was manually annotated as per the rules mentioned in section 4.2.4; however, no lexemes were extracted from it.

In order to have a validation baseline, ZeroR classifier was used first. ZeroR classifier has the simplest classification algorithm the focuses on the target class without using any features. ZeroR guesses the majority class correctly. The majority class is the class that has the highest frequency within a corpus. Although ZeroR is a poor classifier, it is beneficial to determine a baseline for other classifiers.

The ZeroR classifier achieved an F1-measure of 0.37 for AC and 0.38 for NC, which is considered very low when compared to the lowest LB results (0.56) yet this is expected since the corpus contains 5 classes with roughly similar number of records.

After the use of ZeroR and considering its low performance, another baseline needed was considered. An NB classifier was used in six different setups to in effect provide a better baseline. The setups and their results are shown in table 11; all setups outperformed ZeroR as expected.

In the first three setups (first three rows in table 11) 10-fold cross validation was applied for AC, NC, and CNew. The results were higher than the remaining setups where training and testing corpora were different. This was expected because in the first three setups, each corpus was split into training and testing sets, which meant subset of the corpus was used to classify the remaining part of the corpus.

The relatively high results of the last setup may be due to the higher number of training data and to its nature: when AC and NC were used for training, the NB classifier gave higher results when classifying CNew that contains arts and news comments. However, even in that setup, the results were lower than setup three when CNew was used in 10-fold cross validation. NB classification results ranged between 0.446 and 0.626. Given the ease of implementation, the high number of

classes, and the relatively high results when compared to what reported in literature in general, the NB classification results were adopted as a baseline.

Table 12 - NB Classification Results

Setup	Result
AC with 10-fold cross validation	0.547
NC with 10-fold cross validation	0.584
CNew with 10-fold cross validation	0.677
AC for training and NC for Testing	0.46
NC for training and AC for Testing	0.446
AC and NC for training and CNew for Testing	0.626

We now analyse the LB classification results shown in table 13. Four different setups were conducted: AC was classifier using AL and AC, and NC was classified using NL and AL. CNew was not used since AC is foreign to NL and NC is foreign to AL, so there was no need to use a new unseen corpus to validate the LB classification results. However, additional setups will follow in section 6.5

Table 13 - LB Classification Results of Initial Setups

	NL	AL
NC	0.9	0.6
AC	0.6	0.9

The high results achieved when classifying NC using NL and classifying AC using AL were expected since lexemes from the corpus were used to classify the corpus itself. Although this is considered a weak methodology, i.e. to use lexicon extracted from a corpus to classify the corpus itself, it does highlight that order of words in a corpus is important when compared to the 10-fold cross validation followed by NB classifier that disregards the order of words in a corpus during classification.

Primary results contribute to answering the first research question:

1. How to get better understanding of SA of SM comments written in IA?

Results also help addressing the third research objective:

- Compare the performance of an LB classifier with other Machine Learning classifier such as Naïve Bayes (NB) classifier.

In brief, our approach to classify using LB has been shown to be better than NB baseline measures. The most important note was that the LB classifier using AL to classify NC and using NL to classify AC gave better results than the NB using one corpus for training and the other for testing. This may indicate that LB approaches are more dynamic than NB ones in terms of their ability to classify new unseen testing data. Additional setups and analysis will be discussed in section 6.5.

6.2 Analysing the Results of Same-Domain Setups (NC-NL and AC-AL)

The analysis in this section will address the second and third research questions:

- How to improve sentiment classification of SM comments?
- What are the main reasons behind incorrect classification of IA when LB classifier is used?

A comment is considered to be classified incorrectly whenever its label given during manual classification is different from the one given automatically by the classifier. Since there are five different classes (positive, negative, dual, spam, and neutral), five different categories of errors were identified:

- Neutral Errors: Occur when comments are classified manually as neutral and automatically classified as not neutral.
- Negative Errors: Occur when comments are classified manually as negative and automatically classified as not negative.
- Positive Errors: Occur when comments are classified manually as positive and automatically classified as not positive.
- Dual Errors: Occur when comments are classified manually as dual and automatically classified as not dual.
- Spam Errors: Occur when comments are classified manually as spam and automatically classified as not spam.

Table 14 shows percentage of each category in different setups where corpus and lexicon belonged to the same domain. For instance, 31.25% in the first row is the percentage of the

incorrectly classified comments. Only two categories of errors occurred in these two setups. Reasons for incorrect classification follow the table:

Table 14 - Percentage of Errors of Each Category

	AC-AL	NC-NL
Neutral Errors	31.25% (25 comments)	56.52% (39 comments)
Negative Errors	68.75% (55 comments)	42% (29 comments)

6.2.1 Neutral Errors

Inspection of neutral errors identified four different causes, summarized in table 15.

Table 15 - Different Reasons Leading to Neutral Error in AC-AS and NC-NS

Reason		AC-AL	NC-NL
Neutral-R1	Homonyms	40% (10 comments)	10% (4 comments)
Neutral -R2	Presence of Pos lexeme	44% (11 comments)	28% (11 comments)
Neutral -R3	Presence of Pos and Neg lexemes	12% (3 comments)	8% (3 comments)
Neutral -R4	Presence of Neg lexeme	4% (1 comment)	54% (21 comments)

Considering each of these in turn, we examine examples and assess whether alternative NLP tools are able to address the cause.

6.2.1.1 Neutral-R1-Homonyms

Some proper nouns in Arabic have sentimental meaning such as كريم (generous). The automatic classifier will classify such comments as positive although they are neutral. The presence of such lexemes in the comment resulted in incorrectly classifying it as positive instead of neutral. For example, consider the comment below:

كريم نور انت متأكد انو ستار اكاديمي حيتعمل؟ (Kareem Noor are you sure that Star Academy will be active?)

One solution to such a problem would be to detect such proper nouns and exclude them, yet this is not straightforward since such names are too many in Arabic. Plus, it can only be known from context whether the lexicon is expressing a sentiment lexeme or being used as a proper noun. For this purpose, MADAMIRA's NER was used to try to resolve the ambiguity. When the comment above was checked by MADAMIRA's NER, only نور (Noor) was recognized as a proper noun, whereas كريم (Kareem), which was the reason behind incorrect classification, was not recognized as a proper noun but as an adjective (see figure 6.1).

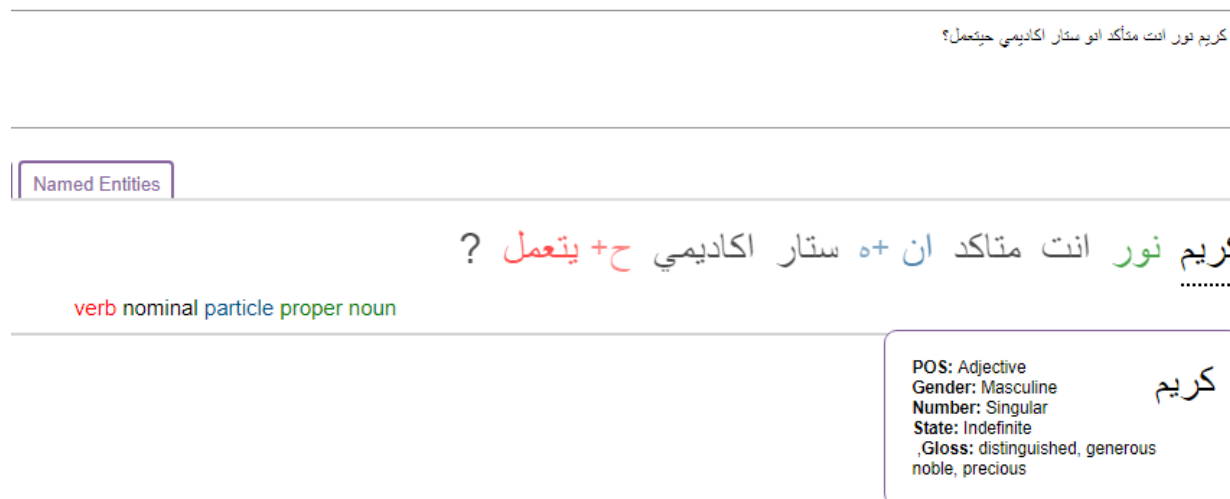


Figure 6.1 - Sample Output from MADAMIRA's NER

Another example is shown below:

اجوي بمهمه عسكريه وانتھو (They came in a military mission then left.)

The word مهمة has two meanings: important and mission. Knowing the exact meaning depends on context. Errors due to homonyms can be caused by many factors; the reason that occurred in this setup was related to proper nouns. Other reasons such as improper tokenization and diacritizations appeared in different setups as will be shown in section 6.2.2.

6.2.1.2 Neutral -R2-Presence of Pos lexeme

Some neutral comments contain positive lexemes in several cases, all related to grammatical aspects.

6.2.1.2.1 Neutral -R2-a-Direct Speech

Such cases occur if the comment is reporting what someone else has said as in the example below:

شيرين عبد الوهاب: النتيجة النهائية من " ذو فويس " عادلة! فخورة بفريد غنام (Shireen Abdul Wahab: The final result of The Voice is Fair, I'm proud of Farid Ghannam)

One solution to this issue would be to search for columns and quotations within the comment. However, given that grammatical rules are not applied in social networks, this may not be a trivial task.

6.2.1.2.2 Neutral -R2-b-Questions

Questions are generally neutral unless they are rhetorical or sarcastic questions. Questions may include positive or negative lexemes without expressing a sentiment, as in the following example:

هل لديك أسئلة تتمنى لو يمكنك توجيهها لمشتري The Voice ؟ (Do you have any questions you wish to ask to the participants of The Voice?) The word تتمنى (wish) is a positive lexeme; however, when it is used in a question, its positive sentiment should be disregarded.

Presence of sentimental lexicon in question is misleading in general, two solutions may help resolve this issue:

Searching for question marks: this may help if it is guaranteed that grammatical rules are applied, which is not the case in Facebook comments, such as the following example:

على فكره مين ربح اليوم فى دربي المغرب الوداد او الرجاء (By the way who won today in Moroccan league, Al Widad or Al Rajaa)

The word ربح (win) exists in a question, yet no question mark is present, and the comment was incorrectly classified as positive.

Searching for question words: This would have been a trivial task if the text is written in Modern Standard Arabic, MSA, where such words are limited and no spelling variants exists such as كيف, لماذا, أين etc. However, in IA, plenty of words exist for each dialect with plenty of possible spelling variants since no spelling rules are applied, and that is where IA POS tagging may provide a proper solution.

6.2.1.2.3 Neutral-R2-c-Informative Speech

Some comments contain sentimental lexemes, but the whole comment is neutral such as:

ياشباب هيك كله حوار بين الموافق بصوت نجم او معارض على صوت نجم (Guys, the whole conversation is about who likes and dislikes the voice of a singer)

The lexeme الموافق (supporter) is positive and the lexeme معارض (opponent) is negative, yet they are not used to express a sentiment.

6.2.1.3 Neutral R3 and R4

The same arguments mentioned for R2 stand for reasons R3 and R4 (presence of positive and negative lexemes) as shown in the example below:

لماذا العرب لم يحل هذه المشكلة (Why didn't Arab solve this problem)

The lexeme المشكلة (problem) is negative; however, it is not used to express a negative sentiment.

Consider presence of negative and positive lexemes in a neutral comment:

صحف غربية: إذا فاز شفيق سيقوم بالإفراج عن المخلوع ومعاونيه (Western newspapers: If Shafic wins, he will release the banished president and his assistants.)

The lexeme فاز (win) is positive, and the lexeme المخلوع (ousted) is negative, yet they are used in an informative comment and not to express sentiment.

6.2.1.4 Neutral Errors and Feature-Sentiment Association

The first two reasons resulting in neutral errors (R1 and R2) can benefit from feature-sentiment association. First, noisy features can be reduced by excluding all terms that do not contribute to the sentiment of the comment such as stop words and other irrelevant words, and secondly, domain-specific lexicon can help resolve the ambiguity of sentiment opinionated lexemes whose that are context dependent. It is worth mentioning here that although AL was extracted from AC and NL was extracted from NC, this does not label them as domain-specific lexicon that is usually more specific in terms of the lexemes it includes and has higher coverage to the terms used in a domain.

Domain-specific lexicons have been shown to be of relevance and can improve performance because they help excluding irrelevant features along with noisy sentiments. Several approaches have been developed: Fahrni and Klenner (2008) studied target-specific sentiment of adjectives. They show how an adjective does not necessarily have a fixed polarity, and that the polarity depends on the noun it is describing. For this purpose, they proposed a model that can reduce sentiment vagueness of adjectives. The first stage of the model was to identify the domain, and then to construct domain-specific lexicons. In their approach, Wikipedia was used for domain detection, and a bootstrapping method was used to construct the lexicon.

Alfrjani et al. (2016) show how different NLP applications such as IR and SA can benefit from domain-specific knowledge. The authors proposed a new semantic model that allows transforming the domain knowledge into a formal ontology. When applied on opinion mining, their approach starts by pre-processing the opinionated textual reviews syntactically and linguistically by applying tokenization, sentence splitting, POS tagging, morphological analysis, and parsing. The second stage in the model requires features extraction or annotation. Finally, a sentiment lexicon is used to decide the sentiment of features.

Wu et al. (2017) studied how different domains use different expressions to express sentiment. They provide an approach that uses different sources to train a domain-specific sentiment classifier. Their approach specifically uses four sources:

- sentiment lexicons
- domain-independent sentiment classifier
- unlabelled data from target domain
- labelled data from target domain

Their approach was tried on Twitter dataset and Amazon product reviews and the result shows that a relatively small number of domain-specific labelled data can improve the classification accuracy.

Given this, our future work includes fine-tuning the sentiment lexicon to become domain-specific, i.e., to split existing lexicon into two: news lexicon and arts lexicon. Where needed, the classification algorithm may be modified to cope for difference in lexicons and domains in addition to giving weights for the lexemes depending on their domain identity. For instance, the lexemes can be given a positive score in one domain and a negative score in another.

6.2.2 Negative Errors

These occur when comment are classified manually as negative and automatically as not negative. Three different reasons were behind this kind of errors as shown in table 16.

Table 16 - Different Reasons Leading to Negative Errors in AC-AS and NC-NS

Reason		AC-AL	NC-NL
Negative-R1	Sarcasm	72.7% (40 comments)	41% (12 comments)
Negative -R2	Misleading Patterns	12.7% (7 comments)	41% (12 comments)
Negative -R3	Negation	15% (8 comments)	17% (5 comments)

6.2.2.1 Negative -R1-Sarcasm

Sarcasm is a form of speech used to ridicule, offend, or belittle an object. It expresses negative attitude toward someone or something and is therefore different than joking whose target is to amuse others. Sarcasm can be expressed directly or indirectly, as we shall see in the examples below. It is usually expressed using positive lexemes and/or negative lexemes. Detecting it is not a trivial task since it is context-dependent and in some cases related to the tone use. Since in our work we are dealing with textual data, the complexity of detecting sarcasm becomes even higher. We have found some traits of sarcasm that can be the starting point of a sarcasm-detecting system that can help reduce the number of comments that are incorrectly classified as dual (since they contain positive lexemes used sarcastically) when they are actually negative. We now describe some traits of sarcasm that were detected in this research:

6.2.2.1.1 Negative -R1-a-Tone-related sarcasm

This usually depends on the way a phrase is said rather than written; the phrase "حلو كتير والله" (very beautiful I swear) in Lebanese accent can either be used as a compliment to describe something that is beautiful, or to express the dissatisfaction. To detect this kind of sarcasm in written texts, the whole thread in which such phrase is used should be checked to see whether they are expressing a positive or negative (sarcastic) sentiment. Since in our work we are dealing with independent textual comments, this task is not considered a priority to use and the positive sentiment will be assumed for such phrases.

6.2.2.1.2 Negative - R1-b-Direct Sarcasm

Using positive lexemes to express aggressiveness or offense. It uses positive and negative lexemes consecutively to express such offense such as:

ألف مبروك الخسارة (Congratulation for the loss)

when the same setups were conducted after adding a space before and after all lexemes, performance decreased by ~7%, so as a trade-off, we will be considering lexemes without adding spaces before or after them.

6.2.2.2.2 Negative - R2-b- Homonyms (Diacritization)

The presence of diacritics can change the meaning of a word. Consider the example below.

ياريت الجزائر وحكومتها تحس على دمها وتسلم المجرمين. (I wish Algeria's government would hand on the criminals). The word تسلم may mean “secure” or “betray” or “hand in” according to the diacritics used. Since IA comments do not use diacritics, the only way to know the meaning of the word is from context. In this specific case, the automatic classifier classified تسلم as a positive lexeme when it was neutral, hence leading to incorrect classification of the comment.

6.2.2.2.3 Negative - R2-c-Valence Shifter

Some regular words act like inverters in that they reverse the sentiment of a comment from positive to negative. Such words, however, have different meanings according to the context in which they are used. Valence shifters have broader scope than inverters, i.e., they do not necessarily flip the polarity entirely. They may reduce the strength of sentiment. However, for the time being, lexemes affected by valence shifters will be considered as negative or positive according to their sentiment without considering the valence shifter. Consider the two examples below:

ينكر الجميل (Yunkir aljameel, To be ungrateful)

انتزاع الرحمة (Intizaa alrahma, To take out mercy)

In the examples above, two positive lexemes were preceded by two words that would result in two phrases of negative sentiments, and although each phrase is already considered a negative lexeme, one word in each phrase is considered positive, so the classifier incorrectly classified the comments as dual instead of negative. Valence shifters are harder to resolve since such words are not few words as the case of inverters where there are only few words such as لا, لم, etc.

To wrap up different cases related to homonyms that appeared in categories Neutral and Negative we can say that homonyms can lead to incorrect classification in four different cases (according the data that we are using, other cases may exist for different data sets):

-Improper tokenization

-Presence of valence shifters

-Presence of diacritization

-Presence of proper nouns

6.2.2.3 Negative-R3-Negation

The presence of inverters may change the polarity of sentimental lexemes. If the automatic classifier is searching for lexemes and did not resolve the presence of inverters well, the comments will be incorrectly classified. On average, 16% of the comments were incorrectly classified because inverters were not considered. In the example, *والله العظيم انو ما صوتك حلو* (which means I swear to God Almighty that you voice is not beautiful), the automatic classifier disregarded the presence of an inverter before a positive lexeme because it did not directly precede it. However, this is not due to the scope of negation used by our approach, but to a very odd phrasing of negation in the first place, at the same time the whole phrase is negative, so according to the automatic classifier, the comment was classified as positive when it should have been classified as negative.

6.2.3 Positive Errors

Positive errors occur when comments are classified manually as positive and automatically as not positive. For example, *نحن نبع الكرامة وعشاق الشهادة والموت لنا غايه* (We are the fountain of honour and lovers of martyrdom and death is an aim to us). The lexeme *والموت* (death) is a negative lexeme, but at the same time the phrase *والموت لنا غايه* (death is an honour to us) is positive, so the automatic classifier classified this comment as dual. Such odd phrases were not common in our corpora and are not likely to appear in high frequency in other corpora because a positive sentiment is expressed using positive lexemes and not negative lexemes.

6.3 Analysing the Results of Cross-Domain Setups (NC-AL and AC-NL)

After classifying each corpus using lexemes extracted from the corpus itself, AC was classified using lexemes extracted from NC and vice versa (CNew was not used since it contains arts and news comments at the same time). F1-measure computed for both setups is almost the same ~56% (see table 13). The relatively high results (higher than NB classification results) may indicate that the lexemes used for classification are domain-independent and that they can be

used to classify unseen corpora. However, due to the relatively small size of corpora, additional experiments are needed before this can be verified. Table 17 shows the percentages of different categories of errors for each setup:

Table 17 - Percentage of Different Errors in AC-NS and NC-AS

Error	AC-NL	NC-AL
Neutral Errors	4.24% (18 comments)	1.19% (5 comments)
Negative Errors	38.82% (165 comments)	36.58% (154 comments)
Positive Errors	23.53% (100 comments)	27.08% (114 comment)
Dual Errors	25.65% (109 comments)	32.78% (138 comments)
Spam Errors	7.76% (33 comments)	2.38% (10 comments)

The reasons behind neutral and negative errors are the same as those discussed in previous sections. As for positive, dual, and spam errors, the reason was that comments contain lexemes that are not in the lexicon, and hence the classifier failed to classify them accordingly. Adding these lexemes to the lexicon will help in partially resolving the issue. A complete lexicon is not a trivial task as the words and phrases that express positive and negative sentiments vary with time and sometimes with contradicting manner. For instance, the word مخيف (which means scary) is recently being used in Lebanese dialect to express positive sentiment. The significant number of missing lexemes can be attributed to cross-domain classification, i.e., using arts lexemes to classify news comments and vice versa. A bigger and domain-specific lexicon may give better results since some lexemes have different sentiments depending on context. For instance, the lexeme “long” is considered positive if used to describe the battery life of a mobile phone and negative if used to describe a lecture or a trip. However, results reported by Baly et al. (2017a) show that there are cases where ignoring the domain and topic gave better results. Below are examples of positive, dual and spam errors.

Positive Error Example:

صوتك ابدآع يارب تفوزى (which means your voice is awesome, we pray to God that you win)

The positive lexeme in the comment (ابدآع) is not found in the lexicon NL, and the comment was hence classified as neutral. As it may be expected, a lexeme used to describe the beauty of something is unlikely to be found in a lexicon extracted from news comment.

Dual Error Example:

الحمد لله أن المولود بصحة أما هي فالله لا يردّها (Thanks to God the baby is fine, as for her, we couldn't care less to whatever happens to her)

The positive lexeme in the comment(الحمد لله) is common to both corpora, however, the negative lexeme (الله لا يردّها) was not found by the automatic classifier and so the comment was automatically classified as positive whereas it was classified manually as dual.

Spam Error Example:

بليبيز تعملو شير لهصفحة (Please share this page)

The example above has a spam lexeme that was not detected, so the comment was classified as neutral instead of spam.

We notice that spam lexemes are almost the same in both corpora. Approximately 56% of the incorrectly classified comments of AC-NL and 62% of those of NC-AL were due to missing lexemes. These numbers were calculated by adding the percentages of positive, dual, and spam errors of table 17. Continuous boosting of the lexicon will resolve errors caused by missing lexemes. The next section addresses the effect of adding lexemes to the lexicon on the classification performance.

6.4 Effect of Increasing the Lexicon Size

Positive, dual, and spam errors discussed in section 6.3 were due to missing lexemes. This section studies the effect of increasing the lexicon size on the classifier's performance.

Let AL-Total represent the union of AL and AL2 (all arts lexemes)

Let NL-Total represent the union of NL and NL2 (all news lexeme).

(See section 4.3.2.1 for more details about AL2 and NL2)

Table 18 shows the results of classifying AC using NL2 and classifying NC using AL2 when compared to the two previous setups. AC-AL2 and NC-NL2 were not tried because AC-AL and NC-NL already gave near perfect results (see table 13) and therefore trying them will only result in poor performance given that AL2 and NL2 are much smaller in size than AL and NL.

Table 18 - Effect of Increasing the Number of Lexemes on Performance of the Classifier

	AL	NL	AL-Total	NL-Total
AC		0.6		0.61
NC	0.6		0.8	

We notice that AC-NL-Total outperformed AC-NS by 5% whereas NC-AL-Total outperformed NC-AS by 24%. This major improvement in the second case is because AL-Total has more lexemes, which supports our claim that increasing the number of lexemes will lead to increase in performance up to a threshold that is yet to be found. Therefore, if the lexemes extracted are treated to generate more lexemes out of them, by adding different suffixes and prefixes (that may refer to different pronouns), and by considering the letters that are used interchangeably such as *o* and *o* and the different variants of the letter *i*, we can improve the recall of the automatic classifier. Moreover, synonyms and antonyms of extracted lexemes can be added to corresponding sets as well. For example, all synonyms of the lexeme “beautiful” can be added to a set of positive lexemes, and all its antonyms can be added to set of negative lexemes. These sets can be also boosted by adding phrases used in different dialects. If a domain-specific classifier is to be built, one that classifies financial news, a corpus of headlines can be prepared and native speakers can be asked to comment on them expressing negative, positive, or dual sentiment to see which keywords are often used for this specific domain. Parallel to that, keywords used in MSA to express a sentiment (related to the domain) can be translated to their dialectal equivalence. In order to check the efficiency of the classifier, the Gold version of the lexicon mentioned in section 4.3.2 was used to classify CNew. The classifier achieved an average F1-measure of 0.54.

6.5 NB Classifier versus LB Classifier

Tables 11 and 19 show classifications results of NB and LB classifiers; table 19 summarizes all LB classification results.

Table 19 - LB Classification Results

Setup	Result
AC- AL	0.92
AC-NL	0.56
NC-NL	0.93
NC-AL	0.56
NC-AL-Total	0.8
AC-NL-Total	0.61
CNew - Gold	0.54

6.5.1 Same-Domain LB Setups versus NB Cross Validation Setups

AC-AL and NC-NL gave much higher results than all NB setups. However, this was expected since sentimental lexemes extracted from the corpora were used to classify the same corpora. Although classifying a corpus using lexemes extracted from the corpus itself is methodologically weak, it does show that its results are much higher than cross validation used by NB classifier, which also uses part of the corpus to classify the remaining parts of the corpus.

6.5.2 Cross-Domain LB setups versus NB Train/Test Setups

Another relevant comparison is to compare the LB setups AC-NL and NC-AL against the NB setups where one corpus is used for training and another corpus is used for testing. This comparison is considered more relevant than the previous one because training and testing data in cases of NB are different from the case when cross validation is used. Concerning LB setups, it will show the performance of an LB classifier when a lexicon extracted from a domain is used to classify a corpus from another domain.

LB classification setups AC-NL-Total and NC-AL-Total outperformed all NB setups conducted on AC and NC. For AC for instance, the lowest LB results was 0.56, and the highest was 0.61, whereas NB achieved 0.46 and 0.547 when different training and testing data were used, and when 10-fold cross validation was used respectively. Only the comments and their manual annotation were input to the NB classifier and no other features were used, which means it was left to the NB classifier to probabilistically determine the lexemes that can represent each class.

On the other hand, the NB classifier outperformed the LB classifier for CNew: the lowest accuracy for NB was 0.626, which is higher than the LB results of classifying CNew using the Gold lexicon.

It is worth mentioning that even the lowest results achieved using LB are considered high when compared to what is reported in the literature for similar tasks such as the SemEval Task 4, especially that our approach is using five different classes and not only three, and the classification becomes harder as the number of classes increases. A direct comparison though is not possible due to the difference in number of classes (2 classes are used in subtasks B and D, and 3 classes are used in subtask E), and because in this work we do not consider the sentiment of a comment relative to the main post as is the case with subtasks B and E. The only subtasks with 5 classes is subtask C, but as mentioned earlier, the 5 classes (positive, highly positive, neutral, negative, highly negative) are different from those adopted in this work (negative, positive, dual, neutral, spam).

6.6 Classification Results of Different Lexicons

Clearly conducting analyses with lexicons developed as part of this research is of value, but it is also important to consider lexicons in general. With this objective in mind three pre-existing lexicons were identified and used to classify NC and AC. The specific lexicons were chosen on the basis of being among the most significant in literature with many research works using them:

1. SIFAAT (Abdul-Mageed and Diab, 2014)
2. NileULex (El-Beltagy, 2016)
3. NRC Emotion lexicon (Mohammad & Turney, 2010; Mohammad & Turney, 2013)

Since the three lexicons mentioned above do not have spam lexemes, two types of setups were conducted:

1-Classify AC and NC using the negative and positive lexemes from the lexicons mentioned above, and the spam lexemes from our lexicon. This setup will be known as With Spam.

2-Exclude the spam posts from AC and NC and then classify them using the three lexicons mentioned above. This setup will be known as Without Spam.

Table 20 shows the results of the setups.

Table 20 - Results of Classifying AC and NC using Different Lexicons

Lexicon Used	AC	NC
AL-Total		0.8
NL-Total	0.61	
SIFAAT With Spam	0.45	0.49
SIFAAT Without Spam	0.34	0.37
NRC EMOLEX With Spam	0.44	0.47
NRC EMOLEX Without Spam	0.33	0.35
NileULex With Spam	0.51	0.61
NileULex Without Spam	0.42	0.53

All setups mentioned in the table 20 benefited from keeping the spam posts and using the spam lexemes. The performance improvement ranged from 8% to 12%. It was also noticeable that NileULex has achieved relatively high results in all setups with one of them (NC with Spam) being higher than those achieved by NC-AL, yet lower than NC-AL-Total. Moreover, it performed better in classifying NC than in classifying AC, probably because it has more negative lexemes than positive ones. As for its high performance compared to the two other lexicons, it is probably due to the nature of its lexemes that are closer in their informal nature to the comments nature than the remaining lexicons, further experimentation needed to fully confirm this.

6.7 Classification Results of Different Corpora

In addition to trying different lexicons to classify our corpora, different setups were conducted to classify different corpora using our Gold lexicon.

Two of the corpora that were used are BBN blog posts corpus, which is a subset of 1200 Arabic (Levantine dialect) sentences chosen from the BBN Arabic-Dialect/English Parallel Text and Syrian tweets corpus consisting of 2000 tweets annotated for sentiment with three classes: positive, negative, or neutral (Salameh et al., 2015). In both corpora, the spam lexemes were disregarded and the results were 0.31 for the Syrian Tweets corpus and 0.36 for the BBN corpus (both numbers refer to average F1-measure). The main reason behind the low performance is the high number of comments that were incorrectly classified as dual. None of the corpora has dual posts (BBN corpus has 1 record manually annotated as dual, but 1 out 1200 is insignificant). Whenever a comment had both positive and negative lexemes, the comment was classified either as negative or positive depending on what class was considered dominant the manual annotators.

Another three-class corpus that was classified using our LB classifier is the Arabic Gold Standard Twitter Data (Refaei & Rieser, 2014). The corpus contains 6514 manually annotated

tweets (negative, neutral, and positive). As per the two previous cases, the main reason behind the low performance is the dual comments. Our LB classifier achieved an average F1-measure of 0.28.

Another corpus that was classified using our lexicon is TAGREED (TGRD) created by Abdul-Mageed et al. (2014), which is a corpus of tweets consisting of 3015 Arabic tweets: 1466 MSA tweets and 1549 dialectal all classified as being mixed, neutral, negative, objective, or positive. TGRD is provided with annotation done by two different human annotators with IAA of 88%. When classifying the corpus using our lexicon, the classification was considered correct whenever it was equal to one of the annotations, and objective was considered to be the same as neutral. The LB classifier achieved an average F1-measure of 0.26.

From the results mentioned above, and when compared to the result of classifying CNew using our Gold lexicon, it was noted that to properly test the performance of an LB classifier and its corresponding lexicon, a corpus that fits original design of the classifier and its lexicon should be used. In our case, five distinct classes are used and three different types of lexemes, with no scores given to intensity of lexemes since the aim is to determine the class. Such constraints limit the ability of the classifier to give high results when used to classify a corpus records for being negative, positive, or neutral only, or as in one of the SemEval Task 4, to distinguish between positive and highly positive tweets.

6.8 Spam Analysis

Spamming refers to sending advertising messages. Although spamming is mainly related to email spam, there are many other media for spam such as instant messages, blogs, and social networking. In our work, we consider a comment to be a spam if it is advertising for a Facebook page, i.e., if it is inviting others to join a page, invitation to watch a movie, or promoting a product, consider the example below:

ممكن نجمع ١٠٠ شخص يحب تحشيش عراقي بليزرز ليك لي بيج (Can we gather 100 people who like Iraqi sarcasm, please like the page)

The comment is encouraging readers to join a page. The lexeme “ليك لي بيج” is the transliteration of “Like the Page”. Some comments may contain spam, positive and/or negative lexemes. However, we found that the spam lexeme is always dominant. 390 spam comments were

analysed, 88% of them contained only spam lexemes, 1% of them had spam and negative lexemes (or spam, positive and negative lexemes at the same time), and 11% of the comments contained spam and positive lexemes. In all of these comments, the presence of negative lexemes in spam comments is insignificant. However, 11% of spam comments contained positive lexemes. This is because spammers tend to use positive lexemes to promote or praise the page they are advertising. The manual taggers extracted 124 distinct lexemes from NC and AC. In NC-NL and AC-AL, the automatic classifier was correct in all cases, which highlights the dominance and efficiency of spam lexemes in detecting spam comments. We then checked the two other setups: NC-AL and AC-NL, we found that in 89% of the cases the classification was correct. This has two possibilities, either many lexemes were common, which turned out to be wrong, or because some of the few common lexemes are found in high frequency, lexemes such as “ممكن لايك” (please like), “WWW” and “YouTube”. This can be used later when weighted lexemes are being used: when every lexeme has a weight according to its accuracy history and frequency of occurrence.

6.9 Negation Analysis

In this section, we analyse different behaviours of inverters to better understand the way they may affect SA.

1-Tokenization: When inverters appear as separate words, they are separated from their target by a space. However, we noticed from our corpora that this is not the case because spelling rules are not followed. Consider the negated lexeme ماتهزأ (don't ridicule). Ideally, a space should separate from the negative lexeme, so when a space was assumed before flipping polarity of lexemes, this phrase will not be properly treated. Unfortunately, improper tokenization is frequent in IA. One solution would be to search of inverters within 0 or 1 space from the target. This will solve the issue of the phrase mentioned above, yet it may ruin other legitimate cases where negation should not be considered: consider the word مشروعة (legal). If we applied the proposed solution mentioned earlier, the classifier would detect روعة (awesome) as a positive lexeme, preceded by an inverter مش, this will lead to incorrect flip of polarity. However, determining this improper tokenization without referring to context would not be possible.

To study this, MADAMIRA's tokenizer was used to tokenize a phrase that should have been tokenized for the LB classifier to operate properly. In the phrase منيحة بس مشروعه (which means

fine but not awesome), it is clear from context that that word مشرعه should have been split into مش روعه . However, when MADAMIRA's tokenizer was tried, the improper tokenization was not recognized, and the adjective at the beginning of the phrase was tagged as a proper noun.

The complexity increases when the inverter used occurs as a prefix. Consider the verb عجه (liked him). This can be negated by adding the letter م (M) to the word. The result will be معجه. However, the same string has a positive meaning, which is admirer or a fan. In addition to that, this kind of inversion will act without a space separating the target from the inverter, so modifying the algorithm to ignore the space between the inverter and the target can be misleading. Another example would be محب. The word may mean lover or did not love at the same time depending on context. Such problems would not occur in MSA because inverters do not appear as prefixes in MSA. Plus, a diacritized text can easily remove ambiguity of such cases.

2-Fake Inverters: The strings representing inverters have other usages not related to negation. Consider the phrase ما أحلاها (how beautiful she is): the phrase consists of a positive lexeme preceded by the same string that is used for negation. For example, ما نحلى (how beautiful) is a positive lexeme that is usually used to praise the beauty of an object, yet this lexeme is not written أ(A) as it should be. We note, however, that in many cases, the targets of these fake inverters consist of four letters, yet this alone is not enough since legitimate negation cases whose targets consisting of four letters also exist. The problem becomes more complex when these same “fake” negation scenarios appear in legitimate negation cases. Consider the phrase بلا احدى صووووووت بلا نيله (not a beautiful voice at all). The same positive lexeme appears preceded by an inverter that is flipping the polarity of the positive lexeme. Another example would be مش احدى صوت (not the most beautiful voice), where the positive lexeme is preceded by an inverter, flipping the polarity of the positive lexeme. Another important observation is that almost all the targets of the fake inverters start with the letter أ(A), but again this alone is not enough since there are plenty of other cases where real inverters flip polarity of lexemes starting with the same letter. One way to reduce the number of misleading cases, is to filter targets consisting of four letters (when pronouns are not used as suffixes such as ما أجملها) and lexemes consisting of all spelling variants of the letter أ such as آ, إ, ؤ, ء, ئ since these are candidate fake targets. Lexemes

not satisfying these conditions are unlikely to be targets of fake inverters whereas those which do can be manually analysed in context and marked as legitimate or fake targets of inverters.

3-Odd Negation: Although real inverters usually flip the polarity of sentimental targets, there are many cases when this is not true. Consider the phrase ما تسب (don't curse): although "curse" is a negative lexeme preceded by a real inverter, the negated phrase itself is still negative. The target in such cases has the same characteristics as other lexemes when negation is valid, i.e., flipping polarity. For example: the phrase ما تزل (don't be sad) has the same POS-features as previous example (both verbs are in present tense), same semantic features (both lexemes are negative), same syntactic features (both lexemes are preceded by the same inverter), and both are expressing orders, yet the overall outcome is different. The modified algorithm mentioned earlier is prone to error because of such cases and resolving it is part of this research's future updates.

4-Implicit Negation: The sentiment of a negated lexeme can be reversed by a dependent clause. Consider the phrase: ما به عيب سوى عبادة الاصنام (he would have been perfect if he didn't worship statues). In other words, the lexeme "perfect" is implicitly negated since the over phrase imply that "he is not perfect." The first part of the comment is positive, but when a neutral phrase was added, the overall sentiment became negative. Such cases are easier to detect in MSA since words to show "exclusion" are limited. By exclusion words we mean words that are used to show how something would have been given a condition, for example, in English we can say "It would have been perfect if it was blue," which means that an object is not perfect yet, but it will be if a certain condition is satisfied. In MSA three common words are used for this purpose لو, إنما, لولا. We illustrate them below with examples:

لو أنه سمع النصيحة, لكان من السعداء (if he had listened to the advice, he would have been happy now)

لولا التعب, لكان العمل ممتعا (work would have been fun if we don't get tired)

لن ينجح سوى المجتهد (only the hard worker will succeed)

لن ينجح إلا المجتهد (only the hard worker will succeed)

إنما طلبت الأخضر (I only asked for the green one)

However, all these cases are not necessarily applicable in IA, where authors can use spelling variants of these words or use them without proper tokenization, or the same lexeme can be used

to express different meaning, such as شو بدك بهاشغلة ولو (why did you interfere in this). In this phrase the word لو is not used to exclude anything. Plenty of these cases exist which makes resolving the issue of exclusion a nontrivial task.

5-Neutral Targets: In addition to their ability to flip the polarity of sentimental targets, inverters may act on neutral targets to produce a sentimental phrase. Consider the example لا صوت ولا صورة (no voice, no picture). The two lexemes “voice” and “picture” are neutral. However, when preceded by the inverter لا(La), the negated phrase will hold a negative sentiment. Detecting such cases is complex because, generally, negating a neutral target results in a neutral phrase. The neutral lexemes mentioned earlier cannot be used by themselves to express a positive sentiment, i.e., saying صوت وصورة is not used a positive phrase. Another example would be مش ناقص, the lexeme ناقص (missing) is considered neutral in IA since it does not express a sentiment as standalone lexeme. However, when preceded by the inverter, the phrase will express as a negative sentiment. Moreover, the same lexeme ناقص can be used as a negative lexeme in MSA as ناقص العقل (brain deficiency) and if preceded by an inverter in MSA, the overall sentiment will be positive.

In summary, negation in IA is not a trivial task, the five cases mentioned earlier are those that appeared in our work, and there may be other cases. The currently identified issues serve as a start work for future research to resolve all aspects of negation.

6.10 Domain Comparison

We noticed that the presence of negative lexemes in neutral comments in NC (54%) is much higher than those of AC (4%), and this is due to the nature of two domains where news usually contains more negative comments and Arts contains more positive lexemes. News usually cover wars, revolutions, economic crisis (so negative lexemes are expected in high frequency) whereas Arts usually mention compliments about artists’ voice, fashion, beauty, etc. Moreover, the occurrence of each category of errors may vary depending on the domain as shown in tables 14 and 17.

It was also noticed that sarcasm was present in higher frequency in Arts comments (49% of incorrectly classified comments were due to sarcasm) than News comments (17% of incorrectly classified comments were due to sarcasm). This is due to the nature of comments in the two

corpora since the Arts corpus contains comments written by fans of different artists where it is frequent to see fans of one artist commenting sarcastically on other artists and their fans.

However, results reported by Alfrjani et al. (2016) show that considering the domain while constructing the lexicon and using domain-knowledge will improve the performance of NLP applications such as sentiment analysis.

Summary

Primary results show that an LB classifier has the potential to classify SM comments written in IA with a good performance. The chapter introduced different categories of errors encountered during classification along with their reasons. It also proposed solutions to some of these categories and then zoomed into some complex cases faced such as negation, sarcasm, and spam. Moreover, it was also found that increasing the number of lexemes in a lexicon improved classification performance and that within-domain lexicon outperformed cross-domain lexicon indicating that a domain-specific lexicon is expected to outperform a general one.

Finally, errors caused by misleading patterns and homonyms may be resolved by an accurate IA part of speech tagger (POS tagger) along with an accurate named entity recognition. The current results achieved when using MADAMIRA's POS tagger and NER showed many issues that did not resolve the issues encountered by the LB classifier.

CHAPTER 7: Conclusion and Future work

7.1 Conclusions

The evolution of the WWW, mobile technology and computers has provided accessible platforms for mass online user interaction. Moreover, the growth of social media has allowed users to post their opinions on diverse objects such as movies, products, policies, and institutions. Posted opinions contain important information to commercial and governmental organizations because they can steer marketing campaigns and help sense the public mood on events such as elections or product launches. However, the huge size and noisy nature of online data make extracting and classifying the sentiment of the comments an infeasible task to be done manually. NLP applications and tools can help in this regard and many different approaches were introduced to address this problem.

Since different languages have differing characteristics the generality of NLP techniques do not always cross language boundaries, and it is fair to say English is the most dominant target language. Specifically, Arabic is one of the languages where resources are scarce when compared to English. Moreover, the morphological complexity and vocabulary richness of Arabic language adds to the difficulty of NLP analysis since tools available for other languages cannot be directly used. Online users tend to use IA, where grammatical and spelling rules are not solid. This hinders processing the text.

The objectives of this work were as follows:

- Investigate (identify) classical techniques used in SA with focus on Arabic language.
- Implement an LB sentiment classifier to classify SM comments written in IA and investigate how it can provide a better understanding of SA of IA.
 - Construct an annotated corpus to be used for SA.
 - Construct an sentiment lexicon
- Compare the performance of an LB classifier with other Machine Learning classifier such as NB classifier.
- Identify main reasons behind incorrect sentiment classifications.

In subsections 7.1.1 through 7.1.4, we shall review these objectives and assess whether the research questions motivating the research have been met.

7.1.1 Investigating Arabic Sentiment Analysis

SA was discussed in sections 2.3 and 2.4 with a focus on social media and Arabic language. The sections highlighted the main advances and showed that many breakthroughs were done in Arabic SA in terms of additional annotated datasets, NLP tools, and sentiment classification. They also discussed the challenges that are facing Arabic NLP and SA. The findings of the investigation were used to adopt the sentiment classification techniques followed in this work and to highlight relevant datasets that can be used. The two main approaches identified in literature for sentiment classification were LB and ML approaches, and both were used in different setups and on different datasets. The literature also helped in choosing the NLP tools that can be used in SA context.

7.1.2 Constructing the Corpus, the Lexicon, and the Classifier

Following the roadmap provided by the literature review, we found that the IA literature can benefit from annotated corpora that address spam, and from a sentiment lexicon. Data collection and usage plan were setup to construct corpora that can be used in SA of IA keeping in mind the main approaches followed in this area and ensuring that data collection and usage were done ethically. The corpora and lexicon annotation highlighted the need of having solid guidelines prior to data collection and annotation to ensure consistency and transparency. Afterwards, an LB classifier was designed that can handle IA. We tried to keep the design dynamic in a way that allows using it for domain classification. However, we have not tried it in this context yet. High-level reports that summarize classification results were also created.

One worth mentioning recommendation is to keep track of IAA and all metadata related to manually classification of corpus and lexicon. Setting a set of clear and written rules for the manual classification will ensure transparency and give more confidence in the classification results. Concerning the lexicon, it is vital to ensure that there is no overlap between entries, i.e., to make sure that the negative and positive lexemes do not have common entries. Moreover, it is worth mentioning that for different languages there may exist some constraints that govern how the lexicon should be constructed. For instance, knowing in advance that for a specific language, there is a set of common phrases used to express positive attitude may help boosting the lexicon.

7.1.3 Comparing LB and NB classifiers

The constructed classifier was tried on the developed corpora and on other external lexicons. The results show that direct comparison of lexicons is not accurate if they do not have the same sets of lexemes: a lexicon containing negative and positive lexemes cannot be directly compared to one that has spam lexemes in addition to positive and negative lexemes. The same thing applies to the classifier itself: a classifier that is designed to classify a corpus containing 5 classes will perform poorly when used to classify a corpus with a different number of classes. The poor classification results of our classifier on different external corpora highlighted this finding.

We started by considering a ZeroR classifier to be our baseline. However, due to the extremely low results of the ZeroR classifier, we adopted NB classification results. Although results were close, the LB classifier outperformed the NB classifier. Moreover, the LB classification results show that the classifier can benefit from using additional features, as there is a room for improving results that were not high enough.

The results also show that an NB classifier classification can benefit from a large and diverse training set as the case of using AC and NC for training and CNew for testing.

Although our classifier uses one feature of regular expressions that can detect repetitions of letter, the results show that spelling inconsistencies are much more complex and need different tools. POS taggers and NER were used at different stages to study whether they can help avoid incorrect classification. Specifically, MADAMIRA was used and has proved that it has great potential in resolving ambiguity in sentiment. However, since our corpora is written in IA, not all the tools gave perfect results, but the results show that improvements in NLP performance will improve sentiment classification as some incorrect classifications were due to incorrect NER for example.

The constructed classifier was also used to classify different external corpora and using external lexicons, and the findings show that there is potential for improving the lexicon and the corpora. One of the improvements would be to add weights and labels to lexemes.

7.1.4 Reasons of Incorrect Classification

Different categories of errors were discussed in chapter 6. The categories show that SA of IA is still a challenging task due to its irregularity. Results also show that NLP tools can help.

Upon completing the first three objectives, the reasons behind incorrect classification (fourth objective) became possible. Some of the findings were expected (such as the effect of negation) and others were not (sarcasm and misleading lexemes). Our analysis showed our negation resolution is primitive and handles only a relatively small number of cases. Odd cases of negation not related to commonly used inverters show that SA could benefit from studying valence shifter in a thorough manner and use them in SA.

One of the error categories, neutral errors, showed that SA could benefit from feature-sentiment associations. Moreover, domain-related lexicon is expected to ensure better classification results.

7.2 Contributions

Briefly, the contribution of this work can be summarized as follows:

1. Preparing resources for IA (corpus and lexicon), with a new class added, spam. Both resources allow testing performance of sentiment classifiers.
2. Using regular expressions to detect letters repetitions that enable resolving one irregular aspect of IA
3. Addressing negation for IA and highlighting different negation cases that needed to be resolved.
4. Categorizing errors and providing different reasons that led to incorrect classification.
5. Implementing a dynamic LB classifier that can be used for SA and domain categorization.
6. Comparing ML classifiers against LB classifiers and highlighting areas of potential improvements in LB approaches.
7. Comparing performance of developed lexicon with external lexicons.
8. Studying how different NLP tools can be used to resolve ambiguity.
9. Discussing the spam class present in FB comments and its effect on SA.

7.3 Future Work

Having done all this work and critically assessed it, there are specific areas that are of interest and relevance to further SA of Arabic SM. The potential areas of future work were detected while developing the corpora and the lexicon and while studying the different categories of errors. Our future work includes the following:

1. Study fake inverters in depth to reduce their negative effect on classification performance. Moreover, the effect on valence shifters, words or phrase that affect the sentiment without inverting it, is another area that needs further study. Results showed that a trivial resolution of negation only handles a small number of negation cases.
2. Construct a corpus of sarcastic comments and propose proper resolutions to sentiment classification of sarcasm. Sarcasm is of special interest to us because of its complex nature and because it is common on social media.
3. Start with the existing lexicon to create domain-specific lexicons. Our findings show that classification results could improve if domain-specific lexicons were used.
4. Transliterated Arabic Sentiment Analysis
 - a. Construct a lexicon for transliterated Arabic, i.e., the Arabic text written in Latin letters.
 - b. Construct a corpus of transliterated social media comments.
 - c. Implement an LB Transliterated Arabic sentiment classifier that uses the two resources mentioned in points one and two.

References

- 10 reasons people use social media. (2013). Retrieved July 11, 2018, from <https://WWW.onepoll.com/10-reasons-people-use-social-media/>
- Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26(3), 12.
- Abdelali, A., Cowie, J., & Soliman, H. (2005, July). Building a modern standard Arabic corpus. In *Proceedings of workshop on computational modeling of lexical acquisition. The split meeting. Croatia, (25-28 July)*.
- Abdelali, A., Darwish, K., Durrani, N., & Mubarak, H. (2016). Farasa: A fast and furious segmenter for Arabic. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (pp. 11-16)*.
- AbdelRaouf, A., Higgins, C. A., Pridmore, T., & Khalil, M. (2010). Building a multi-modal Arabic corpus (MMAC). *International Journal on Document Analysis and Recognition (IJDAR)*, 13(4), 285-302.
- Abdulla, N., Mahyoub, N., Shehab, M., & Al-Ayyoub, M. (2013). Arabic sentiment analysis: Corpus-based and lexicon-based. In *Proceedings of The IEEE conference on Applied Electrical Engineering and Computing Technologies (AEECT)*.
- Abdul-Mageed, M., & Diab, M. T. (2011, June). Subjectivity and sentiment annotation of modern standard arabic newswire. In *Proceedings of the 5th Linguistic Annotation Workshop (pp. 110-118)*. Association for Computational Linguistics.
- Abdul-Mageed, M., & Diab, M. T. (2012, May). AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis. In *LREC (pp. 3907-3914)*.

- Abdul-Mageed, M., & Diab, M. T. (2014). SANA: A Large Scale Multi-Genre, Multi-Dialect Lexicon for Arabic Subjectivity and Sentiment Analysis. In LREC (pp. 1162-1169).
- Abdul-Mageed, M., Diab, M. T., & Korayem, M. (2011, June). Subjectivity and sentiment analysis of modern standard Arabic. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2 (pp. 587-591). Association for Computational Linguistics.
- Abdul-Mageed, M., Diab, M., & Kübler, S. (2014). SAMAR: Subjectivity and sentiment analysis for Arabic social media. *Computer Speech & Language*, 28(1), 20-37.
- Adouane, W., & Johansson, R. (2016). Gulf Arabic Linguistic Resource Building for Sentiment Analysis. In LREC.
- Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In Proceedings of the workshop on languages in social media(pp. 30-38). Association for Computational Linguistics.
- Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *Acm sigmod record* (Vol. 22, No. 2, pp. 207-216). ACM.
- Akra, D. F. (2015). Towards building a corpus for Palestinian dialect (Doctoral dissertation, Birzeit University).
- Alansary, S., Nagi, M., & Adly, N. (2007, December). Building an International Corpus of Arabic (ICA): progress of compilation stage. In 7th international conference on language engineering, Cairo, Egypt (pp. 5-6).
- Alansary, S., Nagi, M., & Adly, N. (2008, December). Towards analyzing the international corpus of Arabic (ICA): Progress of morphological stage. In 8th International Conference on Language Engineering, Egypt (pp. 1-23).

- AlArabiya. (2012). In Facebook [Fan page]. Retrieved ,2012, from
<http://WWW.facebook.com/AlArabiya>
- Al-Ayyoub, M., Gigieh, A., Al-Qwaqenah, A., Al-Kabi, M. N., Talafhah, B., & Alsmadi, I. (2017). Aspect-Based Sentiment Analysis of Arabic Laptop.
- Aleahmad, T., Aleven, V., & Kraut, R. (2009). Creating a corpus of targeted learning resources with a web-based open authoring tool. *IEEE Transactions on Learning Technologies*, 2(1), 3-9.
- Alfrjani, R., Osman, T., & Cosma, G. (2016). A new approach to ontology-based semantic modelling for opinion mining.
- Alfrjani, R., Osman, T., & Cosma, G. (2017, October). Exploiting domain knowledge and public linked data to extract opinions from reviews. In *Knowledge Engineering and Applications (ICKEA), 2017 2nd International Conference on*(pp. 98-102). IEEE.
- Aljamel, A., Osman, T., & Acampora, G. (2015, November). Domain-specific relation extraction: Using distant supervision machine learning. In *Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference on* (Vol. 1, pp. 92-103). IEEE.
- Al-Kabi, M., Al-Ayyoub, M., Alsmadi, I., & Wahsheh, H. (2016). A prototype for a standard arabic sentiment analysis corpus. *Int. Arab J. Inf. Technol.*, 13(1A), 163-170.
- Al-Kabi, M., Gigieh, A., Alsmadi, I., Wahsheh, H., & Haidar, M. (2013). An opinion analysis tool for colloquial and standard Arabic. In *The Fourth International Conference on Information and Communication Systems (ICICS 2013)* (pp. 23-25).

- Almas, Y., & Ahmad, K. (2007, July). A note on extracting ‘sentiments’ in financial news in English, Arabic & Urdu. In *The Second Workshop on Computational Approaches to Arabic Script-based Languages* (pp. 1-12).
- Al-Sabbagh, R., & Girju, R. (2012, May). YADAC: Yet another Dialectal Arabic Corpus. In *LREC* (pp. 2882-2889).
- Al-Sulaiti, L., & Atwell, E. S. (2006). The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics*, 11(2), 135-171.
- Alwakid, G., Osman, T., & Hughes-Roberts, T. (2017). Challenges in Sentiment Analysis for Arabic Social Networks. *Procedia Computer Science*, 117, 89-100.
- Amancio, D. R., Comin, C. H., Casanova, D., Travieso, G., Bruno, O. M., Rodrigues, F. A., & da Fontoura Costa, L. (2014). A systematic comparison of supervised classifiers. *PloS one*, 9(4), e94137.
- Arabic Speaking Internet Users Statistics. (2017). Retrieved July 13, 2018, from <https://WWW.internetworldstats.com/stats19.htm>
- Atserias, J., Attardi, G., Simi, M., & Zaragoza, H. (2010). Active learning for building a corpus of questions for parsing. *Corpus*, 800(11.35), 9-080.
- Badaro, G., Baly, R., Hajj, H., Habash, N., & El-Hajj, W. (2014). A large scale Arabic sentiment lexicon for Arabic opinion mining. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)* (pp. 165-173).
- Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (Vol. 463). New York: ACM press.

- Bahloul, R. B., Elkarwi, M., Haddar, K., & Blache, P. (2014, September). Building an Arabic linguistic resource from a treebank: the case of property grammar. In International Conference on Text, Speech, and Dialogue (pp. 240-246). Springer, Cham.
- Baly, R., Badaro, G., El-Khoury, G., Moukalled, R., Aoun, R., Hajj, H., El-Hajj, W., Habash, N., & Shaban, K. (2017). A characterization study of arabic twitter data with a benchmarking for state-of-the-art opinion mining models. In Proceedings of the Third Arabic Natural Language Processing Workshop (pp. 110-118).
- Baly, R., Badaro, G., Hamdi, A., Moukalled, R., Aoun, R., El-Khoury, G., ... & El-Hajj, W. (2017). Omam at semeval-2017 task 4: Evaluation of english state-of-the-art sentiment analysis models for arabic and a new topic-based model. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 603-610).
- Bamman, D., & Crane, G. (2008, June). Building a dynamic lexicon from a digital library. In Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries (pp. 11-20). ACM.
- Banea, C., Mihalcea, R., & Wiebe, J. (2008, May). A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources. In LREC (Vol. 8, pp. 2-764).
- Baobao, C. H. A. N. G. (2004, December). Chinese-English parallel corpus construction and its application. In Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation (PACLIC 18), at Waseda University in Tokyo.
- Barhan, A., & Shakhomirov, A. (2012). Methods for Sentiment Analysis of Twitter Messages. In 12th Conference of FRUCT Association.

- Baroni, M., & Ueyama, M. (2006). Building general-and special-purpose corpora by web crawling. In Proceedings of the 13th NIJL international symposium, language corpora: Their compilation and application (pp. 31-40).
- Bhuiyan, T., Xu, Y., & Josang, A. (2009, December). State-of-the-art review on opinion mining from online customers' feedback. In Proceedings of the 9th Asia-Pacific Complex Systems Conference (pp. 385-390). Chuo University.
- Bradford, A. (July 24, 2017) Deductive Reasoning vs. Inductive Reasoning. Retrieved July 11, 2018, from <http://WWW.livescience.com/21569-deduction-vs-induction.html>
- Büttcher, S., Clarke, C. L., & Cormack, G. V. (2016). Information retrieval: Implementing and evaluating search engines. Mit Press.
- Carlson, L., Marcu, D., & Okurowski, M. E. (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Current and new directions in discourse and dialogue (pp. 85-112). Springer Netherlands.
- Caruana, R., & Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning (pp. 161-168). ACM.
- Caseli, H. M., Pereira, T. F., Specia, L., Pardo, T. A., Gasperin, C., & Aluísio, S. M. (2009). Building a Brazilian Portuguese parallel corpus of original and simplified texts. In the Proceedings of CICLing.
- Chakrabarti, S., Dom, B. E., & van den Berg, M. H. (2002). U.S. Patent No. 6,418,433. Washington, DC: U.S. Patent and Trademark Office.
- Cho, J., & Garcia-Molina, H. (1999). The evolution of the web and implications for an incremental crawler. Stanford.

- Cho, J., Garcia-Molina, H., & Page, L. (1998). Efficient crawling through URL ordering. *Computer Networks and ISDN Systems*, 30(1-7), 161-172.
- Copyright. (2017). Retrieved July 11, 2018, from <http://WWW.wipo.int/copyright/en/>
- Dave, K., Lawrence, S., & Pennock, D. M. (2003, May). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web* (pp. 519-528). ACM.
- Davis, J., & Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning* (pp. 233-240). ACM
- De Bra, P., & Post, R. D. J. (1994). Information Retrieval in the World-Wide Web: Making Client-Based Searching Feasible. *Computer Networks and ISDN Systems*, 27(2), 183-192.
- DiNucci, D. (1999). Fragmented future. *Print*, 53(4), 32-33.
- Doreswamy, H. K. (2012). Performance Evaluation of Predictive Classifiers for Knowledge Discovery from Engineering Materials Data Sets. *arXiv preprint arXiv:1209.2501*
- Dudovskiy, J. (2017). Deductive Approach (Deductive Reasoning). Retrieved July 11, 2018, from <http://research-methodology.net/research-methodology/research-approach/deductive-approach-2/>
- Dukes, K., & Habash, N. (2010, May). Morphological Annotation of Quranic Arabic. In *LREC*.
- Dzikovska, M. O., Swift, M. D., & Allen, J. F. (2004, May). Building a computational lexicon and ontology with framenet. In *LREC workshop on Building Lexical Resources from Semantically Annotated Corpora*.

- Eirinaki, M., Pissal, S., & Singh, J. (2012). Feature-based opinion mining and ranking. *Journal of Computer and System Sciences*, 78(4), 1175-1184.
- El-Abbadi, N., Khedhair, A. N., & Al-Nasrawi, A. (2013). Build electronic arabic lexicon. arXiv preprint arXiv:1311.6045.
- El-Beltagy, S. R. (2016). NileULex: A Phrase and Word Level Sentiment Lexicon for Egyptian and Modern Standard Arabic. In *LREC*.
- El-Beltagy, S. R., Kalamawy, M. E., & Soliman, A. B. (2017). NileTMRG at SemEval-2017 Task 4: Arabic Sentiment Analysis. arXiv preprint arXiv:1710.08458.
- El-Haj, M., & Koulali, R. (2013). KALIMAT a multipurpose Arabic Corpus. In *Second Workshop on Arabic Corpus Linguistics (WACL-2)* (pp. 22-25).
- El-Haj, M., Kruschwitz, U., & Fox, C. (2015). Creating language resources for under-resourced languages: methodologies, and experiments with Arabic. *Language Resources and Evaluation*, 49(3), 549-580.
- El-Halees, A. (2011). Arabic Opinion Mining Using Combined Classification Approach. In *Proceedings of the International Arab Conference on Information Technology, ACIT (2011)*, Naif Arab University for Security Science (NAUSS), (Riyadh, Saudi Arabia)
- Facebook Company Statistics. (2017). Retrieved July 11, 2018, from <http://WWW.statisticbrain.com/facebook-statistics/>
- Facebook in the Arab Region. (2013). Retrieved July 11, 2018, from <http://WWW.arabsocialmediareport.com/Facebook/LineChart.aspx>
- Fahrni, A., & Klenner, M. (2008, April). Old wine or warm beer: Target-specific sentiment analysis of adjectives. In *Proc. of the Symposium on Affective Language in Human and Machine, AISB* (pp. 60-63).

- Farghaly, A. (2004). A case for an inter-Arabic grammar. Investigating Arabic: current parameters in analysis and leaning. Leiden: Brill, NHEJ, NV Koninklijke, Boekhandel en Drukkerij, 29-50.
- Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4), 14.
- Farra, N., Challita, E., Assi, R. A., & Hajj, H. (2010, December). Sentence-level and document-level sentiment mining for arabic texts. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on* (pp. 1114-1119). IEEE.
- Ferguson, C. A. (1959). Diglossia. *word*, 15(2), 325-340.
- Flacy, M. (2011). "Nearly 300,000 status updates are posted to Facebook every minute". Retrieved July 11, 2018, from <http://news.yahoo.com/nearly-300-000-status-updates-posted-facebook-every-041508056.html>
- Frakes, W. B., & Baeza-Yates, R. (Eds.). (1992). *Information retrieval: Data structures & algorithms* (Vol. 331). Englewood Cliffs, NJ: prentice Hall.
- Goldberg, A. B., & Zhu, X. (2006, June). Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing* (pp. 45-52). Association for Computational Linguistics.
- Green, S., & Manning, C. D. (2010, August). Better Arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 394-402). Association for Computational Linguistics.

- Grossman, D. A., & Frieder, O. (2012). Information retrieval: Algorithms and heuristics (Vol. 15). Springer Science & Business Media.
- Habash, N. Y. (2010). Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1), 1-187.
- Habash, N., Rambow, O., & Kiraz, G. (2005, June). Morphological analysis and generation for Arabic dialects. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages* (pp. 17-24). Association for Computational Linguistics.
- Hajjem, M., Trabelsi, M., & Latiri, C. (2013). Building comparable corpora from social networks. In *BUCC, 7th Workshop on Building and Using Comparable Corpora, LREC, Reykjavik, Iceland*.
- Han, J., Pei, J., & Yin, Y. (2000, May). Mining frequent patterns without candidate generation. In *ACM sigmod record*(Vol. 29, No. 2, pp. 1-12). ACM.
- Haythornthwaite, C. (2005). Social networks and Internet connectivity effects. *Information, Community & Society*, 8(2), 125-147.
- Hamouda, A. E. D. A., & El-taher, F. E. Z. (2013). Sentiment analyser for arabic comments system. *International Journal of Advanced Computer Science and Applications*, 4(3), 99-103.
- Hamouda, S. B., & Akaichi, J. (2013). Social networks' text mining for sentiment classification: The case of Facebook's statuses updates in the 'Arabic Spring' era. *International Journal of Application or Innovation in Engineering & Management (IIAIEM)*, 2(5), 470-478.
- Hangyo, M., Kawahara, D., & Kurohashi, S. (2012, November). Building a Diverse Document Leads Corpus Annotated with Semantic Relations. In *PACLIC* (pp. 535-544).

- Hatzivassiloglou, V., & McKeown, K. R. (1997, July). Predicting the semantic orientation of adjectives. In Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics (pp. 174-181). Association for Computational Linguistics.
- Hatzivassiloglou, V., & Wiebe, J. M. (2000, July). Effects of adjective orientation and gradability on sentence subjectivity. In Proceedings of the 18th conference on Computational linguistics-Volume 1 (pp. 299-305). Association for Computational Linguistics.
- Houngbo, H., & Mercer, R. E. (2014, June). An automated method to build a corpus of rhetorically classified sentences in biomedical texts. In ArgMining@ ACL (pp. 19-23).
- Htait, A., Fournier, S., & Bellot, P. (2017). LSIS at SemEval-2017 Task 4: Using Adapted Sentiment Similarity Seed Words For English and Arabic Tweet Polarity Classification. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 718-722).
- Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.
- Iida, R., & Tokunaga, T. (2014). Building a Corpus of Manually Revised Texts from Discourse Perspective. In LREC (pp. 936-941).
- Itani, M. (2017). Corpus of Arabic social media posts manually classed for sentiment analysis. SHU Research Data Archive (SHURDA). <http://doi.org/10.17032/shu-170008>
- Itani, M., Zantout, R. N., Hamandi, L., & Elkabani, I. (2012, December). Classifying sentiment in arabic social networks: Naive search versus naive bayes. In Advances in

- Computational Tools for Engineering Applications (ACTEA), 2012 2nd International Conference on (pp. 192-197). IEEE.
- Itani, M., Roast, C., & Al-Khayatt, S. (2017a, April). Corpora for sentiment analysis of Arabic text in social media. In Information and Communication Systems (ICICS), 2017 8th International Conference on (pp. 64-69). IEEE.
- Itani, M., Roast, C., & Al-Khayatt, S. (2017b). Developing resources for sentiment analysis of informal Arabic text in social media. *Procedia Computer Science*, 117, 129-136.
- Izwaini, S. (2003, March). Building specialised corpora for translation studies. In Workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives, *Corpus Linguistics*.
- Jia, L., Yu, C., & Meng, W. (2009, November). The effect of negation on sentiment analysis and retrieval effectiveness. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 1827-1830). ACM.
- Jindal, N., & Liu, B. (2006, August). Identifying comparative sentences in text documents. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 244-251). ACM.
- Johnson, C., Shukla, P., & Shukla, S. (2012). On classifying the political sentiment of tweets. *cs.utexas.edu*.
- Kanayama, H., & Nasukawa, T. (2006, July). Fully automatic lexicon expansion for domain-oriented sentiment analysis. In Proceedings of the 2006 conference on empirical methods in natural language processing (pp. 355-363). Association for Computational Linguistics.
- Keshtkar, F. (2011). A computational approach to the analysis and generation of emotion in text (Doctoral dissertation, University of Ottawa (Canada)).

- Khan, K., Baharudin, B. B., & Khan, A. (2010). Automatic Extraction of Features and Opinion-Oriented Sentences from Customer Reviews. *World Academy of Science, Engineering and Technology, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, 4(2), 102-106.
- Kim, S. M., & Hovy, E. (2004, August). Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics* (p. 1367). Association for Computational Linguistics.
- Koehn, P. (2005, September). Europarl: A parallel corpus for statistical machine translation. In *MT summit* (Vol. 5, pp. 79-86).
- Korayem, M., Crandall, D. J., & Abdul-Mageed, M. (2012, December). Subjectivity and Sentiment Analysis of Arabic: A Survey. In *AMLT A* (pp. 128-139).
- Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2), 265-283.
- Liu, Y., Loh, H. T., & Tor, S. B. (2004). Building a document corpus for manufacturing knowledge retrieval.
- Lu, L., Ghoshal, A., & Renals, S. (2013, December). Acoustic data-driven pronunciation lexicon for large vocabulary speech recognition. In *ASRU* (pp. 374-379).
- Lu, Y., Castellanos, M., Dayal, U., & Zhai, C. (2011, March). Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of the 20th international conference on World wide web* (pp. 347-356). ACM.
- Maamouri, M., Bies, A., Buckwalter, T., & Mekki, W. (2004, September). The penn arabic treebank: Building a large-scale annotated arabic corpus. In *NEMLAR conference on Arabic language resources and tools* (Vol. 27, pp. 466-467).

- MBCTheVoice. (2012). In Facebook [Fan page]. Retrieved ,2012, from <http://WWW.facebook.com/MBCTheVoice>
- McCreadie, R., Soboroff, I., Lin, J., Macdonald, C., Ounis, I., & McCullough, D. (2012, August). On building a reusable twitter corpus. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (pp. 1113-1114). ACM.
- Megyesi, B. B., Hein, A. S., & Johanson, E. C. (2006). Building a swedish-turkish parallel corpus. LREC, Genoa, Italy.
- Mining, W. I. D. (2006). Data Mining: Concepts and Techniques. Morgan Kaufmann.
- Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic personalization based on web usage mining. *Communications of the ACM*, 43(8), 142-151.
- Mohammad, S. M., & Turney, P. D. (2010, June). 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 (pp. 26-34). Association for Computational Linguistics.
- Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3), 436-465.
- Monroe, W., Green, S., & Manning, C. D. (2014). Word segmentation of informal Arabic with domain adaptation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (Vol. 2, pp. 206-211).
- Morinaga, S., Yamanishi, K., Tateishi, K., & Fukushima, T. (2002, July). Mining product reputations on the web. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 341-349). ACM.

- Most famous social network sites 2018. (2018). Retrieved July 17, 2018, from <https://WWW.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>
- Mulki, H., Haddad, H., Gridach, M., & Babaoğlu, I. (2017). Tw-StAR at SemEval-2017 Task 4: Sentiment Classification of Arabic Tweets. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 664-669).
- Mustafa, M., & Suleman, H. (2011). Building a multilingual and mixed arabic-english corpus. In Proceedings Arabic Language Technology International Conference (ALTIC) 2011.
- Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2016a). SemEval-2016 task 4: Sentiment analysis in Twitter. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 1-18).
- Nakov, P., Rosenthal, S., Kiritchenko, S., Mohammad, S. M., Kozareva, Z., Ritter, A., Stoyanov, V., & Zhu, X. (2016b). Developing a successful SemEval task in sentiment analysis of Twitter and other social media texts. *Language Resources and Evaluation*, 50(1), 35-65.
- Number of social network users worldwide from 2010 to 2021. (2016). Retrieved July 11, 2018, from <http://WWW.statista.com/statistics/278414/number-of-worldwide-social-network-users/>
- Olteanu, A., Castillo, C., Diaz, F., & Vieweg, S. (2014, June). CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises. In ICWSM.
- Oostdijk, N. (1999). Building a corpus of spoken Dutch. In CLIN.
- O'reilly, T. (2005, September 30). What is Web 2.0. Retrieved July 11, 2018, from <http://WWW.oreilly.com/pub/a/web2/archive/what-is-web-20.html>

- Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In LREC (Vol. 10, No. 2010).
- Pang, B., & Lee, L. (2004, July). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics (p. 271). Association for Computational Linguistics.
- Pang, B., & Lee, L. (2005, June). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the 43rd annual meeting on association for computational linguistics (pp. 115-124). Association for Computational Linguistics.
- Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.
- Pant, G., Srinivasan, P., & Menczer, F. (2004). Crawling the web. In Web Dynamics (pp. 153-177). Springer, Berlin, Heidelberg.
- Pasha, A., Al-Badrashiny, M., Diab, M. T., El Kholly, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., & Roth, R. (2014, May). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In LREC (Vol. 14, pp. 1094-1101).
- Polanyi, L., & Zaenen, A. (2006). Contextual Valence Shifters. Computing Attitude and Affect in Text, 20, 1-10.

- Popescu, A. M., & Etzioni, O. (2007). Extracting product features and opinions from reviews. In Natural language processing and text mining (pp. 9-28). Springer, London.
- Quinlan, J. R. (1987). Simplifying decision trees. *International journal of man-machine studies*, 27(3), 221-234.
- Refaee, E. (Creator), Rieser, V. (Creator). (2014): Arabic Gold Standard Twitter Data for Sentiment Analysis, European Language Resources Association.
- GS2_DataSet_to_HWU(.zip). 10.17861/0e0a6b6b-4892-4c6e-b640-b204e1190cea
- Refaee, E., & Rieser, V. (2014, May). An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis. In LREC (pp. 2268-2273).
- Regular Expressions Info. (2017), Retrieved July 11, 2018, from <http://WWW.regular-expressions.info>
- Riesa, J., Mohit, B., Knight, K., & Marcu, D. (2006). Building an English-Iraqi Arabic machine translation system for spoken utterances with limited resources. In Ninth International Conference on Spoken Language Processing.
- Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46). IBM.
- Roberts, K. (2009, August). Building an annotated textual inference corpus for motion and space. In Proceedings of the 2009 Workshop on Applied Textual Inference (pp. 48-51). Association for Computational Linguistics.
- Rohrdantz, C., Hao, M. C., Dayal, U., Haug, L. E., & Keim, D. A. (2012). Feature-based visual sentiment analysis of text document streams. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2), 26.

- Rosenthal, S., Farra, N., & Nakov, P. (2017). SemEval-2017 task 4: Sentiment analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 502-518).
- Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., & Stoyanov, V. (2015). Semeval-2015 task 10: Sentiment analysis in twitter. In Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)(pp. 451-463).
- Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A., & Perea-Ortega, J. M. (2011). OCA: Opinion corpus for Arabic. Journal of the Association for Information Science and Technology, 62(10), 2045-2054.
- Rytting, C. A., Rodrigues, P., Buckwalter, T., Novak, V., Bills, A., Silbert, N. H., & Madgavkar, M. (2014, June). ArCADE: An Arabic Corpus of Auditory Dictation Errors. In BEA@ ACL (pp. 109-115).
- Saad, M. K., & Ashour, W. (2010, November). Osac: Open source arabic corpora. In 6th ArchEng Int. Symposiums, EEECS (Vol. 10).
- Saif, H., He, Y., & Alani, H. (2012). Semantic sentiment analysis of twitter. The Semantic Web—ISWC 2012, 508-524.
- Salameh, M., Mohammad, S., & Kiritchenko, S. (2015). Sentiment after translation: A case-study on arabic social media posts. In Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies (pp. 767-777).
- Samy, D., Sandoval, A. M., Guirao, J. M., & Alfonseca, E. (2006). Building a Parallel Multilingual Corpus (Arabic-Spanish-English). In Proceedings of the 5th Intl. Conf. on Language Resources and Evaluations, LREC.

- Spiliopoulou, M. (2000). Web usage mining for web site evaluation. *Communications of the ACM*, 43(8), 127-134.
- Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1631-1642).
- Social Networking Statistics. (2013). Retrieved July 11, 2018, from <http://WWW.statisticbrain.com/social-networking-statistics/>
- Somasundaran, S., Ruppenhofer, J., & Wiebe, J. (2008, June). Discourse level opinion relations: An annotation study. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue* (pp. 129-137). Association for Computational Linguistics.
- Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. *Acm Sigkdd Explorations Newsletter*, 1(2), 12-23.
- Statement of Rights and Responsibilities. (2015). Retrieved July 11, 2018, from <https://WWW.facebook.com/terms>
- Tang, D., Wei, F., Qin, B., Zhou, M., & Liu, T. (2014). Building large-scale twitter-specific sentiment lexicon: A representation learning approach. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 172-182).
- Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003, May). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human*

- Language Technology-Volume 1 (pp. 173-180). Association for Computational Linguistics.
- Trusov, M., Bucklin, R. E., & Pauwels, K. (2009). Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site. *Journal of marketing*, 73(5), 90-102.
- Tsunakawa, T., Okazaki, N., & Tsujii, J. I. (2008). Building a bilingual lexicon using phrase-based statistical machine translation via a pivot language. *Coling 2008: Companion volume: Posters*, 127-130.
- Turney, P. D. (2002, July). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 417-424). Association for Computational Linguistics.
- Twitter Developer Documentation. (2017) Retrieved July 11, 2018, from <https://dev.twitter.com/basics/counting-characters>
- What is Intellectual Property.(2017) Retrieved July 11, 2018, from http://WWW.wipo.int/edocs/pubdocs/en/intproperty/450/wipo_pub_450.pdf
- What is Spoken Arabic / the Arabic Dialects?. Retrieved July 11, 2018, from http://WWW.myeasyarabic.com/site/what_is_spoken_arabic.htm
- Wiebe, J., & Riloff, E. (2005, February). Creating subjective and objective sentence classifiers from unannotated texts. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 486-497). Springer, Berlin, Heidelberg.
- Wilson, T., Kozareva, Z., Nakov, P., Rosenthal, S., Stoyanov, V., & Ritter, A. (2013). Sentiment analysis in twitter. In *Proceedings of the International Workshop on Semantic*.

- Wilson, T., Wiebe, J., & Hoffmann, P. (2005, October). Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the conference on human language technology and empirical methods in natural language processing (pp. 347-354). Association for Computational Linguistics.
- World-Wide Web. (2016). Retrieved July 11, 2018, from <http://WWW.ou.edu/research/electron/internet/WWW.htm>
- Wu, F., Huang, Y., & Yuan, Z. (2017). Domain-specific sentiment classification via fusing sentiment knowledge from multiple sources. *Information Fusion*, 35, 26-37.
- Yang, Y., & Liu, X. (1999, August). A re-examination of text categorization methods. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 42-49). ACM.
- Zaghouani, W. (2017). Critical survey of the freely available Arabic corpora. arXiv preprint arXiv:1702.07835.
- Zaidan, O. F., & Callison-Burch, C. (2011, June). The arabic online commentary dataset: an annotated dataset of informal arabic with high dialectal content. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2 (pp. 37-41). Association for Computational Linguistics.
- Zemánek, P. (2001, July). CLARA (Corpus Linguae Arabicae): An Overview. In Proceedings of ACL/EACL Workshop on Arabic Language.
- Zhai, Y., Chen, Y., Hu, X., Li, P., & Wu, X. (2010, October). Extracting Opinion Features in Sentiment Patterns. In Information Networking and Automation (ICINA), 2010 International Conference on (Vol. 1, pp. V1-115). IEEE.

Zhang, K., Narayanan, R., & Choudhary, A. N. (2010). Voice of the Customers: Mining Online Customer Reviews for Product Feature-based Ranking. WOSN, 10, 11-11.

Appendix A: Samples of Data and its Translation

مبروك, (congratulations)

انت احلى صوووووووووت (you have the most beautiful voice)

الف مبروووووووك تستاهل (congratulations you deserve it)

جميل (beautiful)

محببتو (I did not like him)

ما اسمك (what is your name)

مش احلى صوت sot (not the most beautiful voice)

ما أحلاها (how beautiful she is)

روعة (awesome)

ممکن لایک (please like)

Appendix B: Ethics

Excerpt from Facebook's Privacy Policy:

Sharing Your Content and Information

“You own all of the content and information you post on Facebook, and you can control how it is shared through your [privacy](#) and [application settings](#). In addition:

1. For content that is covered by intellectual property rights, like photos and videos (IP content), you specifically give us the following permission, subject to your [privacy](#) and [application settings](#): you grant us a non-exclusive, transferable, sub-licensable, royalty-free, worldwide license to use any IP content that you post on or in connection with Facebook (IP License). This IP License ends when you delete your IP content or your account unless your content has been shared with others, and they have not deleted it.
2. When you delete IP content, it is deleted in a manner similar to emptying the recycle bin on a computer. However, you understand that removed content may persist in backup copies for a reasonable period of time (but will not be available to others).
3. When you use an application, the application may ask for your permission to access your content and information as well as content and information that others have shared with you. We require applications to respect your privacy, and your agreement with that application will control how the application can use, store, and transfer that content and information. (To learn more about Platform, including how you can control what information other people may share with applications, read our Data Policy and Platform Page.)
4. When you publish content or information using the Public setting, it means that you are allowing everyone, including people off of Facebook, to access and use that information, and to associate it with you (i.e., your name and profile picture).
5. We always appreciate your feedback or other suggestions about Facebook, but you understand that we may use your feedback or suggestions without any obligation to compensate you for them (just as you have no obligation to offer them).”

More details about Facebook's data usage policy can be found at the page below:

<https://WWW.facebook.com/policy>

More details about Facebook's terms and conditions are available at the page below:

<https://WWW.facebook.com/terms.php>

Sheffield Hallam University Ethics Form Checklist

The following forms constitute SHU's research ethics checklist:

RESEARCH ETHICS CHECKLIST (SHUREC1)

This form is designed to help staff and postgraduate research students to complete an ethical scrutiny of proposed research. The SHU Research Ethics Policy should be consulted before completing the form.

Answering the questions below will help you decide whether your proposed research requires ethical review by a Faculty Research Ethics Committee (FREC). In cases of uncertainty, members of the FREC can be approached for advice.

Please note: staff based in University central departments should submit to the University Ethics Committee (SHUREC) for review and advice.

The final responsibility for ensuring that ethical research practices are followed rests with the supervisor for student research and with the principal investigator for staff research projects.

Note that students and staff are responsible for making suitable arrangements for keeping data secure and, if relevant, for keeping the identity of participants anonymous. They are also responsible for following SHU guidelines about data encryption and research data management.

The form also enables the University and Faculty to keep a record confirming that research conducted has been subjected to ethical scrutiny.

- For postgraduate research student projects, the form should be completed by the student and counter-signed by the supervisor, and kept as a record showing that ethical scrutiny has occurred. Students should retain a copy for inclusion in their thesis, and staff should keep a copy in the student file.
- For staff research, the form should be completed and kept by the principal investigator.

Please note if it may be necessary to conduct a health and safety risk assessment for the proposed research. Further information can be obtained from the Faculty Safety Co-ordinator.

General Details

Name of principal investigator or postgraduate research student	Maher Muhyiddine Itani
SHU email address	b4046773@my.shu.ac.uk
Name of supervisor (if applicable)	Dr. Chris Roast & Dr. Samir Al-Khayatt
email address	Maher.Itani@hotmail.com
Title of proposed research	SENTIMENT ANALYSIS AND TEXT CATEGORISATION OF DIALECTAL ARABIC TEXT OF SOCIAL NETWORKS
Proposed start date	January 2016
Proposed end date	May 2017
Brief outline of research to include, rationale & aims (500 - 750 words)	Overall Aim To explore, apply and assess sentiment analysis techniques that

Maher

	<p>public social media).</p> <p>Specifically this research programme is to analyse open data from two public groups within Facebook (arts and news). The data was gathered manually respecting Facebook's conditions, in keeping with its copyright conditions (https://www.facebook.com/help/203805466323736). The data is being used as a representative sample of informal online language use – it consists of 2000 public short posts (regarding popular tv shows, etc.) As part of the research conducted before transferring to SHU the data was gathered and 'tagged' to classify statements in terms of the sentiment value (i.e. positive, negative, and neutral). Within my understanding, this respects the ethical considerations highlighted by the Association of Internet Researchers (see Ethical decision-making and Internet research 2.0: Recommendations from the AoIR ethics working committee, 2012). Checking other sources also made us confident that data collection was ethically sound and did not violate any copyrights:</p> <p>http://www.wipo.int/copyright/en/</p> <p>https://www.facebook.com/help/399224883474207</p> <p>and most importantly https://www.facebook.com/terms.php that explicitly mention: "When you publish content or information using the Public setting, it means that you are allowing everyone, including people off of Facebook, to access and use that information". The objective of gathering the data was to provide representative illustrations of informal Arabic. The data was not gathered to understand any more about the individuals who posted to the public pages.</p> <p>The planned work (over 12 months) is to</p> <ul style="list-style-type: none"> report upon the characteristics of the data, and features of it that indicate its sentiment content design, trial and evaluate technical approaches to mechanically classifying the data accurately report the above in the form of a PhD dissertation
Where data is collected from human participants, outline the nature of the data, details of anonymisation, storage and disposal procedures if these are required (300-750 words)	Publicly available textual data were extracted from Facebook and stored in a database, without the contributors names in 2012. The data consist of phrases written in Dialectal Arabic and express common phrases used in everyday life such as "congratulations" and "I don't like this weather"

Maher S

	<p>Traceability: An individual's Facebook name could be identified only if their phrase was unique within the public forum. The research prior to transferring to SHU attributed a sentiment to each post. All public forum posts are already in the public domain and are of a non-sensitive nature. The research while at SHU is to determine if the sentiment class can be computationally predicted. The local working copy had names of authors removed since they are words that are considered useless to the analysis of the text, especially that the phrases themselves are common phrases and do not represent content sensitive to authors.</p> <p>Regarding Human Participation (question 2): We have discussed with the chair of ethics committee whether copying and analyzing public posts constitutes "human participation". If we accept this is "human participation", the data was gathered covertly. However: (i) this is within keeping of Facebook terms and conditions, and (ii) the purpose of research was assessed as unlikely to have any negative impact upon participant (in keeping with Association of Internet Research advice). In support of this the Facebook groups sampled have never required moderation for unacceptable content.</p> <p>Question2 (below) has been completed with respect to the work after transferring to SHU.</p> <p>Data management: Data will be stored on SHU server and a local working copy will be saved on my personal password protected laptop. A data management plan is currently being finalized with the SHURDA staff advice. (This has been delayed to staff illness and vacations - a draft can be provided upon request.)</p> <p>Data collection took place prior to joining SHU. During data collection, data was anonymized and stored on a secure password protected server of Beirut Arab University.</p> <p>After my transfer to SHU my task is to further analyze and study the data collected at Beirut Arab University..</p>
Will the research be conducted with partners & subcontractors?	<p>No</p> <p>(If YES, outline how you will ensure that their ethical policies are consistent with university policy.)</p>

1. Health Related Research involving the NHS or Social Care / Community Care or the Criminal Justice System or with research participants unable to provide informed consent

Question	Yes/No
----------	--------

Maher

1.	Does the research involve? <ul style="list-style-type: none"> • Patients recruited because of their past or present use of the NHS or Social Care • Relatives/carers of patients recruited because of their past or present use of the NHS or Social Care • Access to data, organs or other bodily material of past or present NHS patients • Foetal material and IVF involving NHS patients • The recently dead in NHS premises • Prisoners or others within the criminal justice system recruited for health-related research* • Police, court officials, prisoners or others within the criminal justice system* • Participants who are unable to provide informed consent due to their incapacity even if the project is not health related 	No
2.	Is this a research project as opposed to service evaluation or audit? <i>For NHS definitions please see the following website</i> http://www.nres.nhs.uk/applications/is-your-project-research/	yes

If you have answered **YES** to questions 1 & 2 then you **must** seek the appropriate external approvals from the NHS, Social Care or the National Offender Management Service (NOMS) under their independent Research Governance schemes. Further information is provided below.

NHS <https://www.myresearchproject.org.uk/SignIn.aspx>

* Prison projects may also need National Offender Management Service (NOMS) Approval and Governor's Approval and may need Ministry of Justice approval. Further guidance at: <http://www.hra.nhs.uk/research-community/applying-for-approvals/national-offender-management-service-noms/>

NB FRECs provide Independent Scientific Review for NHS or SC research and initial scrutiny for ethics applications as required for university sponsorship of the research. Applicants can use the NHS proforma and submit this initially to their FREC.

2. Research with Human Participants

Question	Yes/No
1. Does the research involve human participants? This includes surveys, questionnaires, observing behaviour etc. <i>Note If YES, then please answer questions 2 to 10</i> <i>If NO, please go to Section 3</i>	No
2. Will any of the participants be vulnerable? <i>Note 'Vulnerable' people include children and young people, people with learning disabilities, people who may be limited by age or sickness or disability, etc. See definition</i>	N/A
3. Are drugs, placebos or other substances (e.g. food substances, vitamins) to be administered to the study participants or will the study involve invasive, intrusive or potentially harmful procedures of any kind?	N/A
4. Will tissue samples (including blood) be obtained from participants?	N/A
5. Is pain or more than mild discomfort likely to result from the study?	N/A
6. Will the study involve prolonged or repetitive testing?	N/A

Nghia

7	Is there any reasonable and foreseeable risk of physical or emotional harm to any of the participants?	N/A
<i>Note: Harm may be caused by distressing or intrusive interview questions, uncomfortable procedures involving the participant, invasion of privacy, topics relating to highly personal information, topics relating to illegal activity, etc.</i>		
8	Will anyone be taking part without giving their informed consent?	N/A
9	Is it covert research?	N/A
<i>Note: 'Covert research' refers to research that is conducted without the knowledge of participants.</i>		
10	Will the research output allow identification of any individual who has not given their express consent to be identified?	N/A

If you answered **YES only** to question 1, you must complete the box below and submit the signed form to the FREC for registration and scrutiny.

Data Handling

Where data is collected from human participants, outline the nature of the data, details of anonymisation, storage and disposal procedures if these are required (300 -750 words).

If you have answered **YES** to any of the other questions you are **required** to submit a SHUREC2A (or 2B) to the FREC. If you answered **YES** to question 8 and participants cannot provide informed consent due to their incapacity you must obtain the appropriate approvals from the NHS research governance system.

3. Research in Organisations

Question	Yes/No
1 Will the research involve working with/within an organisation (e.g. school, business, charity, museum, government department, international agency, etc.)?	no
2 If you answered YES to question 1, do you have granted access to conduct the research? <i>If YES, students please show evidence to your supervisor. PI should retain safely.</i>	N/A
3 If you answered NO to question 2, is it because: A. you have not yet asked B. you have asked and not yet received an answer C. you have asked and been refused access. <i>Note: You will only be able to start the research when you have been granted access.</i>	N/A

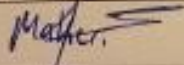
4. Research with Products and Artefacts

Question	Yes/No
1. Will the research involve working with copyrighted documents, films, broadcasts, photographs, artworks, designs, products, programmes, databases, networks, processes, existing datasets or secure data?	yes

M. Jones

2.	If you answered YES to question 1, are the materials you intend to use in the public domain?	Yes
<p>Notes 'In the public domain' does not mean the same thing as 'publicly accessible'.</p> <ul style="list-style-type: none"> Information which is 'in the public domain' is no longer protected by copyright (i.e. copyright has either expired or been waived) and can be used without permission. Information which is 'publicly accessible' (e.g. TV broadcasts, websites, artworks, newspapers) is available for anyone to consult/view. It is still protected by copyright even if there is no copyright notice. In UK law, copyright protection is automatic and does not require a copyright statement, although it is always good practice to provide one. It is necessary to check the terms and conditions of use to find out exactly how the material may be reused etc. <p>If you answered YES to question 1, be aware that you may need to consider other ethics codes. For example, when conducting Internet research, consult the code of the Association of Internet Researchers; for educational research, consult the Code of Ethics of the British Educational Research Association.</p>		
3.	If you answered NO to question 2, do you have explicit permission to use these materials as data? If YES, please show evidence to your supervisor. PI should retain permission.	
4.	If you answered NO to question 3, is it because: A. you have not yet asked permission B. you have asked and not yet received an answer C. you have asked and been refused access. Note You will only be able to start the research when you have been granted permission to use the specified material.	

Adherence to SHU policy and procedures

Personal statement	
I can confirm that:	
<ul style="list-style-type: none"> I have read the Sheffield Hallam University Research Ethics Policy and Procedures I agree to abide by its principles. 	
Student / Researcher/ Principal Investigator (as applicable)	
Name: Maher M. Itani	Date: September 14 th , 2016
Signature: 	
Supervisor or other person giving ethical sign-off	
I can confirm that completion of this form has not identified the need for ethical approval by the FREC or an NHS, Social Care or other external REC. The research will not commence until any approvals required under Sections 3 & 4 have been received.	
Name:	Date:
Signature:	

Name:	Date:
Signature:	

Please ensure the following are included with this form if applicable, tick box to indicate:

	Yes	No	N/A
Research proposal if prepared previously	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Any recruitment materials (e.g. posters, letters, etc.)	<input type="checkbox"/>	<input type="checkbox"/>	X
Participant information sheet	<input type="checkbox"/>	<input type="checkbox"/>	X
Participant consent form	<input type="checkbox"/>	<input type="checkbox"/>	X
Details of measures to be used (e.g. questionnaires, etc.)	<input type="checkbox"/>	<input type="checkbox"/>	X
Outline interview schedule / focus group schedule	<input type="checkbox"/>	<input type="checkbox"/>	X
Debriefing materials	<input type="checkbox"/>	<input type="checkbox"/>	X
Health and Safety Project Safety Plan for Procedures	<input type="checkbox"/>	<input type="checkbox"/>	X
Data Management Plan*	<input type="checkbox"/>	<input type="checkbox"/>	In preparation

If you have not already done so, please send a copy of your Data management Plan to rdm@shu.ac.uk
It will be used to tailor support and make sure enough data storage will be available for your data.

Completed form to be sent to Relevant FREC. Contact details on the website.

Mafer

Snapshot of SHUREC Approval

From: Macaskill, Ann [mailto:sslam2@exchange.shu.ac.uk] **On Behalf Of** Macaskill, Ann
Sent: 10 November 2016 15:07
To: Roast, Chris
Subject: RE: SHUREC1 - CR- MI 2-Sep 9.docx

Dear Chris

The documentation has been scrutinised on behalf of SHUREC and I am satisfied that the data collected by the doctoral student Maher Itani for his research on Sentiment Analysis before he transferred to SHU was collected ethically. It is understood that these circumstances arose because of the absence of a REC in his previous university.

Kindest regards

Ann

Professor Ann Macaskill MA PHD C Psychol PFHEA AFBPS.
Head of Research Ethics/Professor of Health Psychology, Sheffield Hallam University,
Development and Society, Room U0803, Unit 8 Science Park, Sheffield S1 2 WB
Telephone: [+44 \(0\)114 225 4604](tel:+441142254604) Email: a.macaskill@shu.ac.uk
http://www.shu.ac.uk/research/ern/sp_ann_macaskill.html

**Sheffield
Hallam
University**

Appendix C: Research Data Management Policy

Purpose

The University has policies and procedures in place to ensure good research practice and to sustain programmes of excellent and ethical research. Policies are also concerned with research quality promoting the highest standards of integrity, impartiality and respect for data. The University recognises that effective research data management through the research life cycle is a key component of good research conduct and contributes to a culture of research excellence. Research data is a valuable asset and the University supports the principle of open access to research data as set out by the Organisation for Economic Cooperation and Development (OECD) and Research Councils UK (RCUK). Research data refers to any type of data created, collected or generated in a digital or non-digital form that is analysed to produce original research results. The aims of this policy are to:

- support openness and transparency in research undertaken at the University by ensuring research is of the highest integrity and is underpinned by accurate robust data
- promote open access to research data to facilitate data sharing and collaboration and support the University's charitable mission of disseminating research findings
- ensure that the University adheres to the Research Councils UK Common Principles on Data Policy, is compliant with the specific requirements of the EPSRC policy framework on research data and provides accountability for the use of public funds
- establish the responsibilities of researchers in relation to research data management and archiving and set out the University's processes for support and guidance

Policy requirements

1. Data management

1.1. Responsibility for research data generated during a project lies with the principal investigator or in the case of a PhD project, the director of studies. It is their duty to ensure that all members of the research team with access to the research data adhere to good research data management practice. In the case of collaborative projects, if the principal investigator is based elsewhere, the lead researcher at Sheffield Hallam University must take responsibility for all data generated here.

1.2. A data management plan must be produced for all research projects before they commence. Researchers will comply with funder data management requirements. However, where this is not specified, the University will provide a data management plan template for completion.

1.3. A collaboration agreement must be in place with external partners before the start of the research that clearly addresses data management.

2. Live data

2.1. Researchers must comply with funder data management requirements. Where this is not specified researchers must ensure that all active research data is stored securely on the University networked storage system in both original and processed formats. The University has created a central research data

file store (the SHU Research Store) for this purpose and will provide advice on technical solutions for research data storage and archiving. Metadata describing the structure and content of the data must be regularly created and updated for project continuity purposes. If research data needs to be stored temporarily on portable storage devices, such as laptops in the field or cloud storage, the researcher must ensure that this is done securely and that they comply with the University's policies on electronic data encryption.

3. Archiving

3.1. Primary research data produced by University researchers that underpin a publication, which are of potential long-term value and/or support a patent application, must be stored centrally and published when possible to ensure good research practice at the University.

3.2. Primary research data, whether in digital or hard copy, may be archived in the SHU Research Data Archive or in an external research data repository. Data must be stored for a period at least as long as that required by any funder or sponsor of the research, any publisher of the research or as set out in the University's Research and Knowledge Transfer Records Retention Schedule.

3.3. It is considered good practice to archive all data in a format that will guarantee long-term access and with sufficient metadata to aid discovery to encourage follow-up research. Researchers must also comply with specific funder data management requirements.

3.4. All data that are retained must be registered with the SHU Research Data Archive, whether they are hosted by the University or maintained elsewhere, even if access to the data is restricted.

4. Open access

4.1. Researchers must be aware of, and comply with, their funders' requirements for data management including archiving and sharing. If applicable, data must be prepared and offered for deposit in an open access data repository within the timeframe stipulated by the funder unless there are valid reasons not to do so. The latter could include commercial confidentiality, infringement of intellectual property rights, contractual agreements, ethical, legal or regulatory obligations, or where the cost of doing so would be prohibitive.

4.2. Even if the funder of the research does not require it, researchers are encouraged to make their archived data accessible to others close to the publication date of any research outputs relying on the data. The data should be in citeable form. This supports the integrity of the University's research and will be beneficial for the research community.

4.3. Exclusive rights to re-use or publish research data should not be handed over to commercial publishers or agents without retaining the rights to make the data openly available for re-use unless this is a condition of funding.

4.4. Published research outputs reporting publicly funded research must include a short statement describing how and on what terms any supporting research data may be accessed. Research outputs deposited in SHURA should also include this statement.

5. Re-using third-party data

5.1. Researchers that gain access to and use open research data, or any data generated by others, must do so in a manner that respects the contexts under which it was created and must adhere to the same frameworks and observe any restrictions that may have been imposed during data collection.

5.2. All users of research data must formally cite the data they use. The obligation to recognise through citation and acknowledgement the original creators of the data must be respected in all cases.

6. Support and further information

6.1. The University will provide guidelines, advice and training on research data management, including data management plans, costing of research data management into research proposals, storage and data protection, creation of descriptive metadata, intellectual property and Freedom of Information requests for all researchers.

7. Scope

7.1. This policy applies to all publicly-funded research, whether internally or externally funded, and is considered best-practice for all other research.

7.2. Contractual obligations from an external funder or sponsor of the research will take precedence over the stipulations in this policy.

This policy was last updated in January 2017.

Appendix D: Data Management Plan

DMP title

Project Name My plan (SHU Template)

Principal Investigator / Researcher Maher Itani

Institution Sheffield Hallam University

Data Collection

What data will you collect or create?

No data will be collected; the research consists of analysing data that have been collected prior to joining SHU.

How will the data be collected or created?

N/A

Documentation and metadata

What documentation and metadata will accompany the data?

Data consists mainly of an excel file containing 2000 records and 2 fields: each records consist of a phrase written in dialectal Arabic and described using one of five specific labels. The five labels (negative, positive, spam, neutral, or dual) represent the sentiment of each record as specified by the student conducting the research.

Ethics and Legal Compliance

How will you manage any ethical issues?

Research is using data that is not protected by copyrights; the data being used consists of 2000 Facebook comments posted publicly on public pages and no on users' personal profiles. The comments consist of short phrases that do not constitute any artistic or scientific work. Moreover, all these comments where written in dialectal Arabic and contain expressions used in everyday life such as "congratulations", "the weather is nice", etc.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

I checked what data can be copyrighted, the data I'm operating on are not copyrighted material since they do no constitute any genuine work of any kind, moreover, the mechanism in which they were posted (selecting public audience and posting them on a public page) make them available for the public. Details can be found at <https://WWW.facebook.com/help/203805466323736>.

Storage and Backup

How will the data be stored and backed up during the research?

We will be using SHU's server to store the data, a local working copy is kept on a personal laptop that is password restricted. The laptop is kept in a physically secure place all of the time and not shared.

How will you manage access and security?

Selection and Preservation

What data are of long-term value and should be retained, shared, and / or preserved?

All data (raw and analysed) will be deposited in the University's Research Data (SHURDA) before the end of the research project. The data will be retained in the

archive for a period of 10 years since the last time any third party has requested access to the data. When depositing the data, no further changes to data formatting will be required as all necessary actions will have been conducted as the research progresses

What is the long-term preservation plan for the dataset?

All raw data will be made available.

Data Sharing

How will you share the data?

Are any restrictions on data sharing required?

We will deposit and share our data at the end of the project without any delay. Any research outputs that are published will contain a statement that refers to the underlying datasets and how these datasets can be accessed; any restrictions to access will be outlined and justified in this statement. The raw anonymized data and the data collection methodologies will be made available on a Creative Commons with Attribution (CC-BY) or equivalent license. supervisory team.

Responsibility and Resources

Who will be responsible for data management?

SHU

What resources will you require to deliver your plan?

Resources available at SHU

Appendix E: Classifier Design

Filter by:

Inverters.csv Negs.csv Poss.csv Posts.csv Spams.csv

Figure E.1 - Form Used to Upload Lexicon, Inverters, and Comments to be Classified

☒ Add ☐ Empty First

☒ Add ☐ Empty First

☒ Add ☐ Empty First

☒ Add ☐ Empty First

☒ Add ☐ Empty First

Drag a column header here to group by that column	
Type	Count
Inverters Count	14
Negs Count	1275
Poss Count	1015
Posts Count	1000
Spams Count	100

Figure E.2 – Form Used to Load Corpus, Inverters, and Lexicon

Start from:

of records to process:

Clear old result? ☐

Figure E.3 – Form Used to Specify Number of Records

Drag a column header here to group by that column									
Post	Classification	Spams	Spams Count	Positives	#Positives	Negatives	#Negatives	#FlippedToPos	#FlippedToNeg
الله معاك يا ابو سمره	POSITIVE		0	ابو سمره الله معاك	2		0	0	0
انت الاروع يا حاتم و مبروك النجاح مسبقا وموفق ياذن الله	POSITIVE		0	الاروع النجاح ياذن الله مبروك	4		0	0	0
قيلعلى صوتك يا ابو سمره	POSITIVE		0	ابو سمره قيلعلى صوتك	2		0	0	0
لا تنسى تبلغ نجاننا لشعب العراق من اخوانهم العراقيين	POSITIVE		0	اخوانهم نجاننا	2		0	0	0
مش عارفه شو عاجبها فيه	NEGATIVE		0		0	مش عارفه شو عاجبها	1	0	0

Figure E.4 – Sample of Classification Results

Classification ▾				
	Post	Spams	Spams Count	Positives
▸	Classification: DUAL			
▸	Classification: NEGATIVE			
▸	Classification: NEUTRAL			
▸	Classification: POSITIVE			
▸	Classification: SPAM			

Figure E.5- Sample of Grouping Results

Drag a column header here to group by that column				
Post	Classification	Spams	Spams Count	Positives
				مبروك
مبروك يا فريد	POSITIVE			0 مبروك
مراد مبروك	POSITIVE			0 مبروك
الف مبروك مراد	POSITIVE			0 مبروك

Figure E.6 – Sample of Filtering Results

Classification	Spam	Original Positive	Flipped To Positive	All Positive	Original Negative	Flipped To Negative	All Negative
DUAL	0	40	0	40	32	1	33
NEGATIVE	0	0	0	0	19	0	19
NEUTRAL	0	0	0	0	0	0	0
POSITIVE	0	100	0	100	0	0	0
SPAM	5	0	0	0	0	0	0

Figure E.7- Sample of Classification Summary

Classification	Count	Total Posts	Percentage
DUAL	17	101	17.0
NEGATIVE	13	101	13.0
NEUTRAL	5	101	5.0
POSITIVE	61	101	60.0
SPAM	5	101	5.0

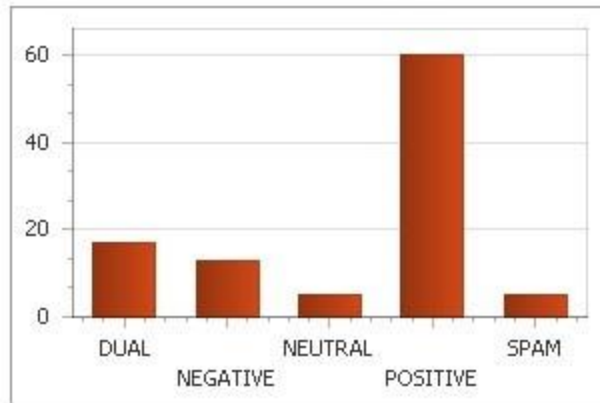


Figure E.8- Frequencies and Percentages of comments of each Class

Pattern	Count
مبوك	14
احلى	12
شهه	3
منافقين	2
احلا صوت	2
ما شاء الله	2
ملك الطرب	2
عش	2
بالتوفيق	2
كذابين	2
كلنا معاك	2
ابو سمره	2
احسن	2
الاروع	2

Figure E.9 – Summary Showing Frequency of Lexemes