LINKING ARABIC SOCIAL MEDIA BASED ON SIMILARITY AND SENTIMENT

A thesis submitted to the University of Manchester for the degree of

Doctor of Philosophy

In the Faculty of Engineering and Physical Sciences

2016

By

Samah Alhazmi

School of Computer Science

CONTENTS

Word Count: 45,754

LIST OF TABLES

LIST OF FIGURES

LINKING ARABIC SOCIAL MEDIA BASED ON SIMILARITY AND SENTIMENT

Samah Alhazmi

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy, 2016

ABSTRACT

A large proportion of World Wide Web (WWW) users treat it as a social medium, i.e. many of them use the WWW to express and communicate their opinions. Economic value or utility can be created if these utterances, reactions, or feedback are extracted from various social media platforms and their content analysed. Some of these benefits are related to e-commerce, marketing, product improvements, improving machine learning algorithms etc. Moreover, establishing links between different social media platforms, based on shared topics and content, could provide access to the comments of users of different platforms. However, studies to date have generally tackled the area of content extraction from each type of social media in isolation. There is a lack of research of some aspects of social media, namely, linking the references from a blog post, for example, to information related to the same issue on Twitter. In addition, while studies have been carried out on various languages, there has been little investigation into social media in the Arabic language. This thesis tackles opinion mining and sentiment analysis of Arabic language social media, particularly in blogs and Twitter. The thesis focuses on Arabic language technology blogs in order to identify the expressed sentiments and then to link an issue within a blog post to relevant tweets in Twitter. This was done by assessing the similarity of content and measuring the sentiments scores. In order to extract the required data, text-mining techniques were used to build up corpora of the raw blog data in Modern Standard Arabic (MSA) and to build tools and lexicons required for this research. The results obtained through this research contribute to the field of computer science by furthering the employment of text-mining techniques, thus improving the process of information retrieval and knowledge accumulation. Moreover, the study developed new approaches to working with Arabic opinion mining and the domain of sentiment analysis.

DECLARATION

No portion of the work referred to in the dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

COPYRIGHT

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property University IP Policy (see http://documents.manchester.ac.uk/display.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University's policy on Presentation of Theses.

ACKNOWLEDGEMENTS

Undertaking this PhD has been a truly life changing experience for me, and it would not have been possible to do without the support and guidance that I received from many people.

I would like to first say a very big thank you to my supervisor Mr. John McNaught for all the support and encouragement he gave me during my PhD. Without his guidance and constant feedback this PhD would not have been achievable.

Many thanks also to Dr. William Black and Dr. Yannis Korkontzelos for their help. I gratefully acknowledge the NaCTeM team for providing encouragement.

To my husband, who has been by my side throughout this PhD, living every single minute of it, and without whom I would not have had the courage to embark on this journey in the first place.

To my mother, for her kindness, and for being with me every time I needed her help with my children.

To my father, for his encouragement and for making me strong.

To my lovely sisters, for being helpful.

PUBLICATIONS

Alhazmi, S., & McNaught, J. (2013). Generating an Arabic Sentiment Corpus from Social Media. *Lancaster Workshop on Arabic Corpus Linguistics*, 56–59.

Alhazmi, S., Black, W., & McNaught, J. (2013). Arabic SentiWordNet in Relation to SentiWordNet 3.0. *International Journal of Computational Linguistics, 4*(1), 1–11.

ACRONYMS AND ABBREVIATIONS

TechTerms	Technologies Terms
ArTerMine	Arabic TerMine
WN	WordNet
SWN	SentiWordNet
ASWN	Arabic SentiWordNet
NB	Naïve Bayes
KDD	Knowledge Discovery in Databases
CRM	Customer Relationship Management
ML	Machine Learning
RSS	Rich Site Summary
NLP	Natural Language Processing
WWW	The World Wide Web
MSA	Modern Standard Arabic
IR	Information Retrieval
ER	Entity Recognition
IE	Information Extraction
SO	Sentiment Orientation
XMI	XML Metadata Interchange
ACE	Automatic Content Extraction

CHAPTER 1: INTRODUCTION

This chapter presents an overview of the problem that this research investigates and presents the motivation for this study. Furthermore, it outlines the research aim and objectives of the study.

1.1 Motivation and research problem

This research is motivated by how people interact using social media, and by the difficulties and challenges in gaining access to information about other users' feelings and opinions on a given subject in any social medium, for example, in a blog. According to Feldman et al. (1998), these communication changes have resulted in several challenges in the field of opinion mining and sentiment analysis; they also present several opportunities. From an economic point of view, the decision by most businesses to go online has its own consequences. People express their opinions and feelings about the products and services that such organisations offer on various social media, e.g., through blogs and on Twitter[1], or on various e-commerce sites. These sites have prompted the companies to seek new marketing strategies; they also present an opportunity for companies to get to know the needs and problems of their consumers, with the result that they are better able to gain trust for their brands.

As Feldman et al. (1998) and Balahur and Balahur (2009) have highlighted, research in the market analysis and the text mining sectors has shown that the emergence of micro-blogging[2] has changed and affected the decisions made by individuals and companies; these decisions are often related to opinions—and rumours—found in blogs.

[1] www.twitter.com

[2] Micro-blogging is a broadcast medium that exists in the form of blogging. A microblog differs from a traditional blog in that its content is typically smaller in both actual and aggregated file size.

According to Binali et al. (2009), opinion on the Internet is similar to a form of virtual currency that can determine whether or not a product sells in the market. Other studies show that the information available in news articles has a strong influence on society, since opinions on aspects of 'social phenomena' are readily available on blogs, forums and review sites (Conrad and Schilder, 2007; Pak and Paroubek, 2010). Many of the companies' functions would be easier to perform, given the availability of such online opinions. This is in line with the strategies which the organisation could take in its advertising, its 'business intelligence' and its 'competitive vigilance' (Binali et al., 2009).

As stated by Pak and Paroubek (2010), new online social systems make it easier to gather the opinions of writers from, for example, Twitter and blogs; this is vital to organisations in pursuit of a competitive advantage. Moreover, if the information on these blogs is readily available, then the companies can determine the market segments in which their products sell well, and as a result improve their influence on the clients.

Balahur and Balahur (2009) consider that sentiment analysis and opinion mining of the Web for information present a way to gather information about the opinions of consumers so that the organisations can act in a way that enables them to gain consumers' support. Strapparava and Mihalcea (2008) contend that opinion mining and sentiment analysis can aid in decision making by companies through the identification of new ideas and the ability to find solutions to their technological and economic problems.

According to Pang et al. (2002), the opinions found on the Web in various online social media have resulted in changes to how several communities interact with each other and in the 'elaboration' of laws and policies. In the past, communities interacted with one another through time-consuming questionnaires; now, with the increase in the use of technology, the use of opinions expressed on social media has become more widespread.

In addition, Pang et al. (2002) claim that it is important for such information to be readily available to Internet users. For instance, data about communities' laws and policies is found on micro-blogs, blogs, and other social networks.

A lack of information in social media about the markets and other data, or the unavailability of information, implies that use of social media such as Twitter has often led to slow decision-making processes in organisations, and that results in financial loss (Wilson et al., 2005). Pang and Lee (2008) state that the lack of information for policy makers in various communities could lead to poor decisions that might, in the end, have a negative impact on the concerned Web-based social community's interaction.

According to Melville et al. (2009), in social networks, people often talk about subjects or matters and give their views and emotions about them where it would be hard to express them in public; the result is that social media is often the best place to learn the facts or the most common opinions about a given subject.

As Agrawal and Siddiqui (2009) assert, other barriers such as the family could be the problem in talking about such matters in public; these might include psychological problems, educational problems and other concerns. The Web provides a 'hidden identity' that encourages freedom of expression without fear. For these reasons, and despite the importance of opinion mining and sentiment analysis, there are challenges in finding out peoples' true feelings and emotions on given matters.

1.2 Domain

Using methods drawn from text mining research—opinion mining and sentiment analysis in this case—data was analysed and categorised into different types. The method of analysis used in the research could be applied to several datasets. For instance, our research methods involved the use of a sequence of data to gather useful sentences, phrases and words; the same method was then applied to all datasets in order to infer and

show similarities or differences between datasets. Data provides a starting point for study and analysis, from which researchers can draw, test and analyse hypotheses.

Many studies in opinion mining and sentiment analysis have used reviews e.g. movie reviews, as their datasets, in both English and Arabic (Hu and Liu, 2004; Prabowo and Thelwall, 2009; Rushdi-Saleh et al., 2011; Turney, 2002). For the purpose of our research, to generate a new corpus for Arabic sentiment analysis, the data were gathered from different blogs about technology (Alhazmi and McNaught, 2013). The use of blogs on technology provided various challenges, as will be explained in Chapter 4, because of the various ways of writing blog posts.

The reason for choosing Arabic as a case study is because little research has been done for this particular language (Abbasi et al., 2008a; Abdul-Mageed et al., 2011; Al-Subaihin et al., 2011; Farra et al., 2010); this research aims also at studying the tools currently available and making them supportive of the Arabic language.

1.3 Research aim, objectives, questions and hypotheses

1.3.1 Research aim

Our research aimed to investigate Arabic social media, using Arabic technology blogs and Twitter as a case study. Criteria for inclusion of data in the study were primarily based on information content – by this we mean a blog post that includes any information related to technology, e.g. features of a type of mobile phone, software utilities, etc. – and various opinions about the same information – from Twitter.

The research also involved comparisons of the different sentiment values expressed in the opinions conveyed through social media. The purpose of this research was to find out whether it would be possible to extract information from Web pages that reflect the emotions of the writer.

The Arabic language was used as a case study in our research; however, our framework is flexible and constructed so as to be applicable to other languages.

1.3.2 Research objectives

More precisely, the objectives and associated activities of our research as shown in Figure 1.1 are as follows:

Figure 1.1. *Research objectives and activities*

Specific research objectives are:

RO1 Design the research framework, as shown in Figure 1.2.

RO2 **Resources and tools:**

RO2.1 Research Corpus: generate an Arabic corpus from Arabic technology blogs (Alhazmi and McNaught, 2013). Then, annotate and evaluate the datasets based on three main tasks; by assigning these tasks to a small group of Arabic annotators to find the following: TechTerms (technology terms), Facts (Neutral attributes and factual descriptions) and Sentiment (Positive or Negative opinions).

RO2.2 Build a TechTerms list: as we are trying to make it an open list, this list should consist of:

- English technological terms, e.g. iPhone, Galaxy S3, etc., with all possible transliterations to Arabic (for example, Google was transliterated as qwql قوقل, jwjl جوجل, gwgl غوغل).
- English names of technology companies e.g. Samsung, Apple, etc., with all possible transliterations to Arabic.
- Arabic technological terms, e.g. (mgrdwn مغردون).
- Arabic names of technology companies, e.g. ($rkp HAswb شركة حاسوب).

RO2.3 Adapt the existing TerMine[3] tool (Frantzi et al., 2000) to support the Arabic language (ArTerMine).

RO2.4 Construct the Arabic SentiWordNet (ASWN) (Alhazmi et al., 2013) in relation to the English version of SentiWordNet 3.0 (Esuli and Sebastiani, 2006).

RO3 Link blog posts to the relevant tweets: by dealing with each blog post separately, and then collecting all retrievable tweets that are relevant to the same information in the blog post. Finally, evaluate the sentiments expressed in both, and also rank the sentiments expressed in the tweets according to whether they are positive, negative or neutral.

1.3.3 Research questions

This research addressed the following main questions:

RQ1 To what extent can we advance the state of the art in opinion mining to search for, identify and classify sentiments expressed in micro-blogs about the content and opinions of blogs?

[3]http://www.nactem.ac.uk/software/termine/

RQ2 To what extent can we automatically, and in the absence of world knowledge, detect emotion, as a component of sentiment, in both blogs and micro-blogs, where this is expressed implicitly?

1.3.4 Research hypotheses

Below is an outline of fundamental research hypotheses that were assessed and analysed throughout the research process:

RH1 It should be feasible to link content in different types of social media such as blogs and micro-blogs by using text mining techniques, using the Arabic language as a case study.

RH2 Analysis of implicitly expressed emotions can improve sentiment-based techniques in linking different types of social media, using the Arabic language as a case study.

RH3 Involving *multi-words-terms*[4] in a hybrid clustering method should enhance the quality of the outcome clusters.

1.4 Contributions of this thesis

This research makes the following contributions:

RC1 Linking of two types of social media (blogs and Twitter) by analysing contents and sentiments.

RC2 Generation of a research corpus using blogs about technology written in Arabic (Alhazmi and McNaught, 2013).

[4] The use of Arabic multi-words terms (MWTs) has played an important role in this research in two phases (Phase 2 and Phase 3) of the research framework shown in Figure 1.2. Further explanation is in Chapters 5, 7 and 8.

RC3 Implementation of the ASWN with relation to the existing English version of the SentiWordNet 3.0. Taking into account the sentiment scores for each word in the ASWN (Alhazmi et al., 2013: 8).

RC4 ArTerMine: adapting the existing TerMine tool to be supportive of the Arabic language for the extraction of useful terms from a corpus.

RC5 Building a TechTerms list that covers all technological terms in our corpus, and adding more terms to make an extensible resource.

RC6 A hybrid method used for generating clusters from our corpus. This method combined three levels of clustering (1) using the raw text in a blog post only as the first level to yield clusters of single words; (2) using multi-word terms (MWTs) related to this blog post only as the second level to yield clusters of MWTs; (3) applying both together as the third level to yield clusters including both single words and MWTs. It is important to mention that in this thesis, a "cluster" means a group of words and/or multi-word phrases, extracted from a text such as a blog post; and "clustering" means extracting significant words and phrases from a text.

1.5 Thesis structure

The remaining chapters of this thesis are as follows:

Chapter 2 This chapter presents a literature review of opinion mining and sentiment analysis and shows the importance of this domain in different languages. We consider how opinion mining and sentiment analysis capabilities are configured into applications that can play an important role in social networks through social media.

Chapter 3 In this chapter, we give an overview of our research framework which has four phases and outline the workflow of our data through these phases.

Chapter 4 In this chapter, a summary of characteristics of the Arabic language in its written forms that are relevant to text mining operations. Then, the design of our corpus generated from Arabic technology blogs (RC2) will be described, highlighting the challenges and the difficulties faced through the data preparation and annotation stages, and the approach taken to evaluating this corpus.

Chapter 5 In this chapter, the roles of distinct levels of analysis in text mining are discussed, before concentrating on providing improvements to lexicons and tools for Arabic sentiment analysis. Due to the lack of a sentiment lexicon for Arabic, we have built the ASWN as an Arabic text mining research resource (RC3). Similarly, we have adapted the ArTermine tool (RC4) for extracting Arabic terms from the TerMine in its English version. Finally, to satisfy the need for a small lexicon that contains technologies terms, we have made an open list (RC5).

Chapter 6 This chapter describes in detail Phase 1 of our research framework: corpus data processing. This was carried out by using U-Compare and then applying the tagged data to the ArTerMine.

Chapter 7 This chapter focuses on Phase 2 of our research framework: clustering of blog posts and identification of similarities. A hybrid clustering method was used in this step; this is considered as one of our research contributions (RC6). This was done by combining and integrating three levels of clustering (1) using the raw text in a blog post only as the first level to yield clusters of single words; (2) using multi-word terms

(MWTs) related to this blog post only as the second level to yield clusters of MWTs; (3) applying both together as the third level to yield clusters including both single words and MWTs.

Chapter 8 In this chapter, we mention in detail – Phase 3: analysis of Twitter data – in our research framework. After reviewing some additional background work on the mining of Twitter data, we show how multi-word terms are used as search keywords in the Twitter API to collect tweets related to each blog post. Through this phase, we made use also of an 'Arabic converter' that can change the transliterated text to a readable format of Arabic. Finally, we discuss the quantity of collected tweets regarding each blog post.

Chapter 9 This chapter covers in detail one of our research contributions (RC_1), Phase 4 of this research framework. Some background work is mentioned at the beginning of this chapter; this is followed by a report of our experiment. We took the last collection of useful clusters for each blog post—the output of Phase 2—together with the related tweets collected for this particular blog post—the output of Phase 3—and used these as input for Phase 4; by this, we mean that our experiment in this phase was applied to each blog post separately. The ASWN and the TechTerms list were applied at the lexicon look-up stage to determine words related to sentiments and to technology. Next, the collection of useful clusters and related tweets was applied at the next stage to classify sentiments in both. Finally, the sentiments were ranked, and the final evaluation was carried out.

Chapter 10 In this chapter, we provide a summary of this thesis, followed by a discussion of the extent to which the work supports our research

hypotheses and research questions. We conclude with a discussion of future work indicated by our findings.

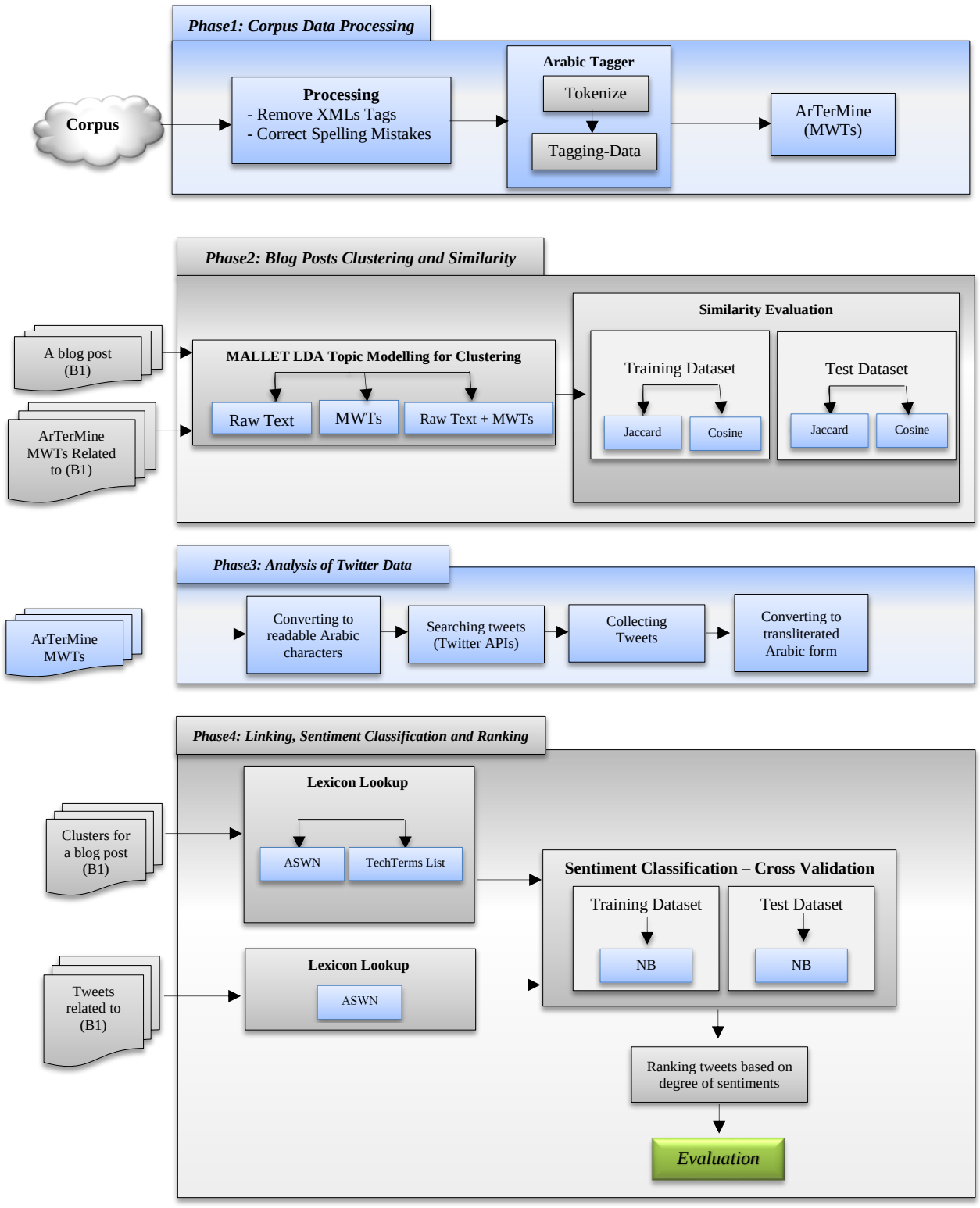


Figure 1.2. *Research framework*

CHAPTER 2: LITERATURE REVIEW

In this chapter, we review the previous work and technical approaches to opinion mining and sentiment analysis, starting from more general information about this domain and narrowing down to our specific interest (blogs and Twitter) in social media. The scope of the literature review is shown in Figure 2.1.

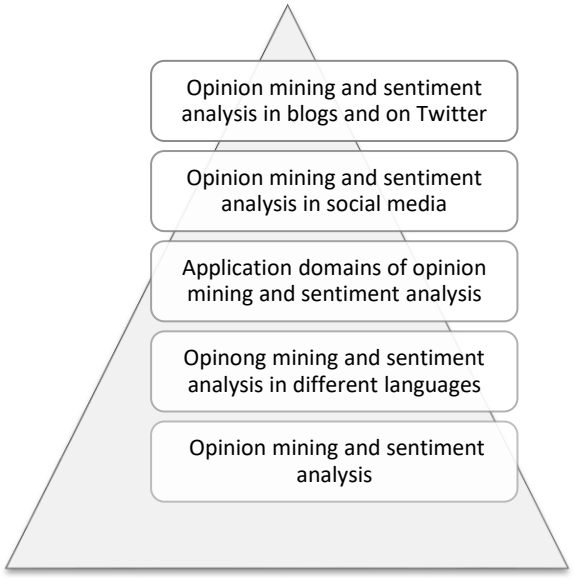


Figure 2.1. *Literature review*

2.1 Overview

This research aimed to assess the emotional state of people communicating online in Arabic through blogs, and to draw a connection between the opinions expressed in these blogs and those expressed on the Twitter social network. This entire process is about knowledge discovery, or 'mining'. According to Feldman et al. (1998) knowledge discovery in databases (KDD) pays attention to the automated analysis of an extremely large amount of data as well as the detection of useful patterns found within this data. As

the majority of the tasks regarding KDD are related to structured databases, there has been a considerable amount of effort required in managing the vast amount of data that is accessible.

Customarily, databases store data in the shape of structured information records and offer techniques for querying so as to extract all records whose details match the user's query. More recently, though, discoveries in KDD have offered an innovative group of techniques and tools intended for discovering useful information in databases. The objective of such research is frequently known as data mining—the main field includes text mining and opinion mining—and has been described by Feldman et al. (1998) as 'the nontrivial mining of hidden facts' which were previously unidentified.

Various methods in these categories entail the implementation of statistical investigation and machine-learning methods in order to enable the automatic discovery of information and data patterns in databases, and to offer user-guided environments intended for the efficient discovery of data; this occurs in spite of the huge amount of Web-based data that seems to be simply in sets of unstructured text (Farra et al., 2010; Feldman et al., 1998).

2.2 Opinion mining and sentiment analysis

Figure 2.2 presents a general framework for opinion mining and sentiment analysis (Lo and Potdar, 2009). This outlines the various processes used in mining opinions, classifying the sentiment(s) expressed in items and features, and measuring the strength of the sentiment(s).

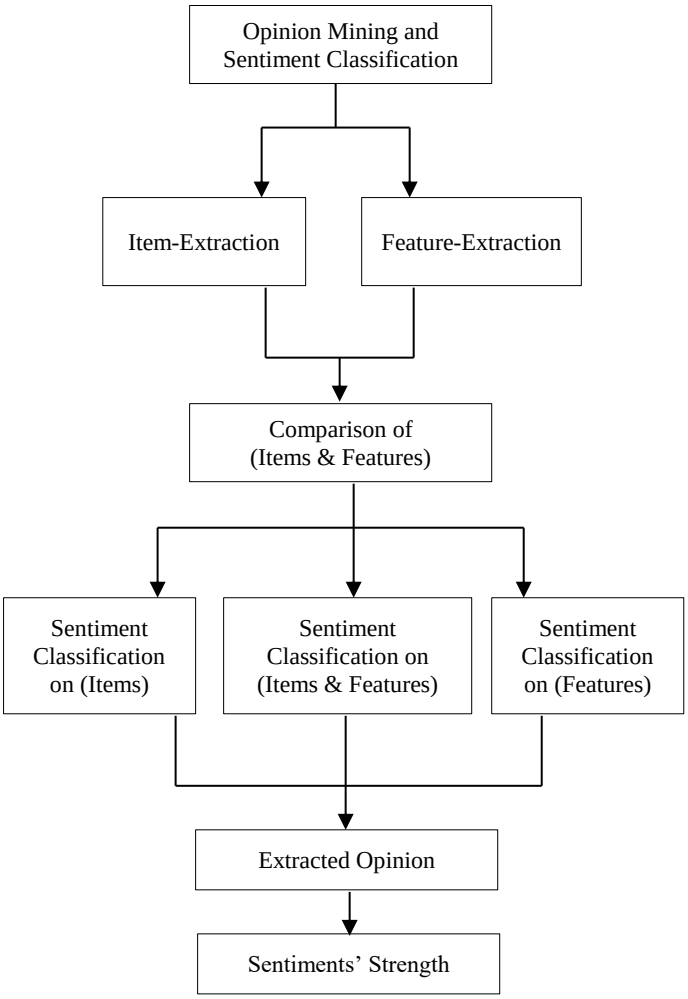


Figure 2.2. *General framework for opinion mining and sentiment analysis (Lo and Potdar, 2009)*

2.2.1 Opinion mining variants

'Opinion mining' is generally considered to be a wide-ranging term that involves a combined judgment on a number of issues and aspects aimed at finding a common understanding (Liu, 2015). According to Strapparava and Mihalcea (2008), the term 'feeling' (in this thesis, the term 'feeling' and the term 'sentiment' used are interchangeably.), regarding opinion mining, is described as the awareness of subjective

knowledge of emotion. This is a significant description that shows that feeling is a small component of emotion, indicating the subjective knowledge process.

In contrast, Dave et al. (2003) state that there are diverse kinds of views and that not all opinions are subjective, nor do all opinions have a sentiment linked to them. In this situation, an objective opinion could be recognised as one of a professional kind.

Opinion mining is an assessment-based task that first appeared in research carried out by Dave et al. (2003), who described it as a specified set of techniques applied to assessment-based texts that express opinions regarding an object, which can be an organisation, a person or a product. In this situation, opinion mining is intended to extract components and attributes of the object that have been stated in a document, and also to decide whether the given comments are positive, negative or neutral.

In real life, people often defer to other people's opinions regarding any decision-making procedure. Given the recent appearance of Web-based sources intended for the expression of opinion, these areas are now under intensive research. Large numbers of people are currently expressing judgments on the Web. Therefore, to utilise these opinions for decision making, it is necessary to use automated management and processing of text in order to recognise the opinion(s) being articulated in the document. For example, according to Agrawal and Siddiqui (2009), businesses make use of social media investigation and monitoring to produce customer insight reports for better assessment of their desired shopping behaviour.

Sentiment polarity analysis is used to recognise negative and positive opinions demonstrated in any given document. At present, a number of research projects are looking at the documents level (Farra et al., 2010). For example, blog-based documents are categorised as positive or negative, depending on the opinion articulated in them. This procedure demands phrase/sentence-level processing.

Agrawal and Siddiqui (2009) state that a classic method is to utilise a lexicon of positive and negative words in making a decision. However, making use of the method of pre-tagging words by fixing the polarity is insufficient. The polarity of a given sentence cannot be assessed by merely looking at the occurrence of words depending on the framework in which they happen. Frequently, sentiments expressed with similar words, for example, *"short", "long", "shiny", "rough"*, and so on, can convey differing opinions depending on the context (Agrawal and Siddiqui, 2009).

Nevertheless, it is proposed that nouns and verbs are able to stand as powerful signs of sentiment. According to Devitt and Ahmad (2007a) and Devitt and Ahmad (2007b), the employment of SentiWordNet[5] is useful in discovering sentiment polarity in business financial news. This method is analogous to the fundamental arrangement employed in the assessment of polarity relations of citations by Piao et al. (2007).

Esuli and Sebastiani (2005) state that opinion mining is a new paradigm at the forefront of digital information retrieval (IR), a computational linguistics tool that extracts information not by means of the topics of various documents but by means of the opinions expressed in the documents. Opinion mining has a vast set of systems such as Customer Relationship Management (CRM) for assessing and tracking users' opinions regarding products or political candidates, such as those articulated in many social and collaborative online forums.

2.2.2 Opinion mining and sentiment analysis in different languages

According to Denecke (2008), sentiment analysis in a multilingual situation presents numerous challenges. Statistical methods and lexical methods require training data based on lexical as well as linguistic resources. Producing these resources is an extremely time-consuming activity because it frequently involves manual work. There are essentially two

[5] http://sentiwordnet.isti.cnr.it/

methods that are applicable in the situation of multilingual sentiment analysis. In this scenario, Denecke (2008) assessed a corpus-based method plus a lexicon-based method for multilingual subjectivity analysis. In the lexicon-based method, target language subjectivity is categorised based on the translation of an available lexicon. The corpus-supported methods offer subjectivity interpretation of the corpus intended for the target language. A statistical categorising application was trialled in the consequential corpus (Denecke, 2008; Liu et al., 2005).

In a scenario of sentiment classification within a multilingual arrangement, Denecke (2008) followed local grammar[6] methods which were intended for sentiment classification inside a multilingual structure for languages like Arabic, English and Chinese. Local grammar methods in this sense rely on establishing frequent collocates of collocates in a domain-specific corpus. This involves choosing domain-related keywords while evaluating the distribution of words in a domain related to a specific document to determine how this compared to the distribution of words in a common language corpus. This type of approach is due to Gross (1997), based on Harris's earlier work in relation to sublanguages (1991). Figure 2.3 shows the processing pipeline for sentiment analysis.

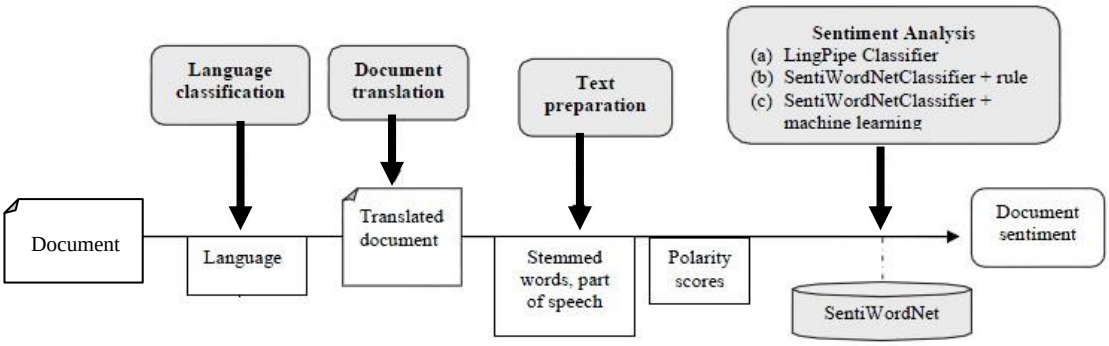


Figure 2.3. *Processing pipeline for sentiment analysis (Denecke, 2008)*

Furthermore, sentiment analysis is also used in different languages on the Web; it is often employed as a means of bringing about the transformation of data, information

[6] For more information about local grammar method, see Ahmad et al. (2005).

and opinions. For assessment of Web-based information and data in social forums, we make use of sentiment analysis (Thelwall et al., 2010).

Figure 2.4 shows a framework for opinion mining done by Binali et al. (2009), which describes the process of extraction of several desired features from the text through text mining methods.

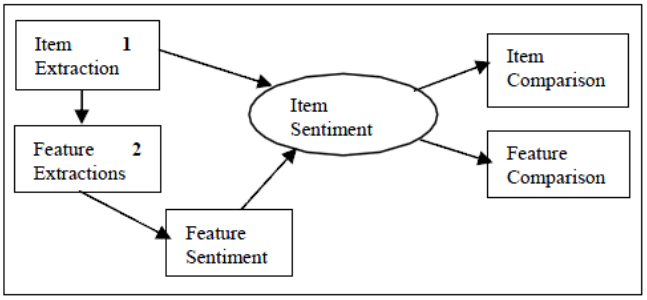


Figure 2.4. *Opinion mining framework (Binali et al., 2009)*

The use of Arabic and English content stylistic and syntactic characteristics in the sentiment-classification dataset is very useful and effective for the overall extraction process. These characteristics are more common and appropriate across languages. For example, lexical, syntactic and structural characteristics have been productively employed in stylo-metric testing studies carried out using the Chinese, English, Greek and Arabic languages (Abbasi et al., 2008a; Thelwall et al., 2010). Figure 2.5 shows an approach towards extracting Arabic features.

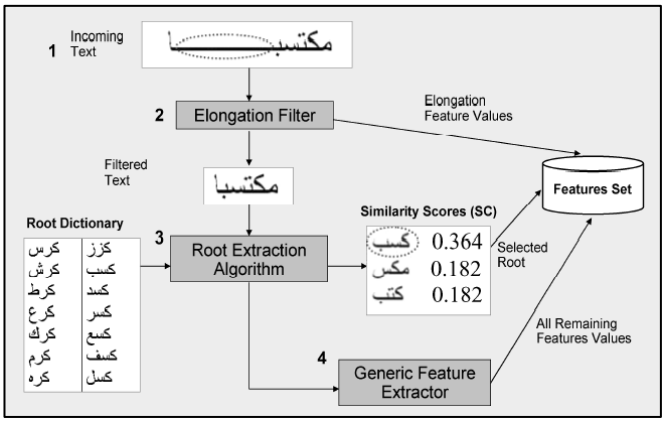


Figure 2.5. *A method of extracting Arabic features (Abbasi et al., 2008a)*

Furthermore, Table 2.1 shows another approach that can be used for the extraction of Arabic text and English text. This table shows us a simple transformation method that we can use for the possible extraction of Arabic text emotional features (Almas and Ahmad, 2007).

Table 2.1. *Metaphors used in English and Arabic financial reporting (Almas and Ahmad, 2007)*

Source	Positive		Negative	
	English	Arabic	English	Arabic
Physics/ Movement	accelerate	*tasaro'*, تسارع	slowdown	*tabatoa*, تباطؤ
	up	*fawq*, فوق	down	*tahat*, تحت
	rise	*ertifa'*, ارتفاع	fall	*enkifaad*, انخفاض
	ascent	*so'od*, صعود	descent	*hoboot*, هبوط
Biologi-cal	healthy	*sihay*, صحي	sick	*mareed*, مريض
	strong	*qawi*, قوي	Weak	*dh'eef*, ضعيف
	grow	*nomow*, نمو	stagnate	*rookood*, ركود
	vital	*hayawi*, حيوي	puny	*hazeel*, هزيل
Other	snowball	*korat thalj*, كرة ثلج	bubble	*foqa'a*, فقاعة
	boom	*tafra*, طفرة	bust	*enhiyar*, انهيار

In the application of sentiment analysis for the Arabic language, one of the main fundamental processes is 'feature extraction'. Because of the challenging morphological features of the Arabic language, it is extremely hard to extract enhanced features for data analysis; the best and most common way of doing this is through the use of root extraction algorithms (Abbasi et al., 2008a; Conrad and Schilder, 2007).

According to the research and experimental working of Abbasi et al. (2008b), sentiment analysis is used for Web-based social networks. In this scenario, Takamura et al. (2004) have assessed the efficiency of their characteristics and tested feature collection methods intended for sentiment categorisation of English and Arabic forums.

The Arabic language is a Semitic language known for its morphological and syntactic complexity. According to Tounsi and Genabith (2010) parsing Arabic sentences is a hard job, often because of the comparatively free word order of the Arabic language.

Many useful tools have been developed for the English language; however, the area of sentiment analysis of Web information written in Arabic is yet to be well established.

As Al-Subaihin et al. (2011) state that the Arabic language is composed of formal and informal expressions. In Arabic-speaking communities it is normal to engage in a sociolinguistic phenomenon known as diglossia, or code-switching, between languages or language varieties; in other words, the linguistic standards employed in the media are different from the spoken language employed in everyday life. Arabic websites are mostly written in the casual/informal variant of the Arabic language (Al-Subaihin et al., 2011).

2.2.3 Application domains of opinion mining and sentiment analysis

At present, opinion mining and sentiment analysis have applications in a variety of everyday fields and are used for various purposes. Here a number of applications for opinion mining are outlined:

- ***Shopping***

  Possibly the best-known area of opinion mining is the corporate analysis of customer shopping behaviour, which is utilised as a means of providing customer service support for consumers. According to Binali et al. (2009), presently customers are aggressively involved in comparing shopping experiences on the Web. Well-known websites such as Amazon (www.amazon.com) permit customers to state their opinions about purchases they have made on their websites.

- ***Entertainment***

  Entertainment is also one of the main areas in which opinion mining can be useful. Through the application of movies and home TV, viewers are able to rapidly read views and opinions on current releases as well as on well-known TV programs and movies. At present, there is the huge Web-based IMDb (Internet Movie Database), which provides online reviews of movies and TV programmes. Sites such as this act

as a support for entertainment in that they provide a guide for people who are uncertain about what to watch (Binali et al., 2009; Conrad and Schilder, 2007; Melville et al., 2009; Vechtomova, 2010).

- ***Government***

  Governments are making use of opinion mining. Opinion mining applications are used to determine current opinions on public policy; as such, they offer election candidates a great deal of support and information concerning the particulars of opinion polls. This information is very useful in helping politicians to recognise their areas of strength and weakness relative to the desires of the electorate (Mitchell, 2012; Binali et al., 2009).

- ***Research and development***

  Opinion mining supports research and development tasks. Here, product reviews can be employed by manufacturing corporations to enhance the characteristics of a product and offer a platform for innovation. An online system could present platforms for consumers to design products and to submit their designs to manufacturers (Binali et al., 2009; Prabowo and Thelwall, 2009).

- ***Marketing***

  Businesses are currently making use of opinion mining for corporate marketing. For this purpose, businesses are taking note of client opinions to improve the development and delivery of customer products and to offer a great deal of enhanced support for customer needs management (Binali et al., 2009; Leskovec et al., 2007; Orimaye, 2011).

- ***Education***

  Opinion mining is also playing a key role in e-learning systems. Here users' opinions are employed to assess academic institutions and teaching faculty (Binali et al., 2009; Stavrianou and Chauchat, 2008).

2.3 Opinion mining and sentiment analysis in social networks

2.3.1 Online social networks communication

This research classifies, filters and retrieves information regarding online communication. According to Abbasi et al. (2008a), the Arabic language is more 'complex' in syntax and morphology than English and therefore needs more processing so that it can be used in opinion mining and sentiment analysis. Thus, through the stemming process, the affixes of the word are removed after pre-processing (Agrawal and Siddiqui, 2009).

Previous research has explored ways to make information available to users on the Internet (Duwairi and Alshboul, 2015; Esuli and Sebastiani, 2005; Melville et al., 2009). Research carried out by Al-Subaihin et al. (2011) shows that because of the increase in the number of people who use blogs and other social networks, there is a need for opinion mining and sentiment analysis. This research describes new methods to be used in information retrieval on the Internet.

2.3.2 Opinion mining in the social Web

The expansion of Web technology has occurred through the combined expansion and development of Web 2.0, also known as the 'Social Web'. As suggested by Balahur and Balahur (2009), Web 2.0 has been the cause of novel and interesting Web-based social trends. Conversely, the ability to articulate opinions 'anywhere, by anybody or anything', in online social network-based forums, blogs, review websites, and so on, has made it feasible for people all around the globe to make a more enhanced and better informed decision at the time of purchasing services and products.

Similarly, businesses and public servants are more knowledgeable about the influence they have on people because the huge number of opinions articulated provides unbiased and worldwide feedback directly to them. From another perspective, the

uncontrolled expression of opinions has offered a means to create aggressive messages, leading to anti-social behaviour and other negative actions. Because of the huge volume of information disseminated, computerised systems have to be developed to tackle these issues (Balahur and Balahur, 2009).

The majority of work on sentiment analysis has been performed on extremely subjective text types, such as product and movie reviews, and personal blogs (Conrad and Schilder, 2007; Duwairi and Alshboul, 2015; Duwairi et al., 2015; Melville et al., 2009; Vechtomova, 2010).

A variety of research has ensured that there are systems for the analysis of English and other languages, but there has been limited research on the Arabic language. Text in a document needs to pass through several phases, including conversion, removal of words that do not make sense, stemming, feature selection, construction of a vector, feature weighting, classifier construction, and evaluation of the classifier (Abbasi et al., 2008a, Chen and Salem, 2008; Abdul-Mageed et al., 2011; Conrad and Schilder, 2007).

As Dave et al. (2003) point out, there has been a problem in the past in getting information that has not been ‘structured’. A lack of stored information and poor methods of querying the database have been noted in the literature. However, the knowledge of the researcher would enable the extraction of information through new techniques if such data were available. This can include machine learning, even though this is very challenging when it comes to the Arabic language (Al-Subaihin et al., 2011; Dave et al., 2003). The automatic detection of sentiments of different viewers of information can be helpful. The objective of such a system is to come up with efficient ways in which organisations can get immediate feedback from customers and determine how they feel about their products in an automatic manner (Dave et al., 2003).

2.3.3 Motivation for opinion mining research in social networks

At the financial level, the globalisation of marketplaces united by means of a situation whereby people are able to openly express their viewpoints on various products or corporation on blogs, RSS[7] Feed, forums or e-commerce websites has led to a transformation in the marketing policies of businesses. In this scenario, Pang and Lee (2008) state that the augmentation of responsiveness to client requirements and criticisms has brought about a heightened awareness of brand reputation and trust.

New technology-based experts in market analysis, as well as in fields like Natural Language Processing (NLP), have verified that given the recently developed opinion phenomena, choices intended for economic action are not specified simply with factual information; they are also considerably influenced by negative opinions and false rumours (Pang and Lee, 2008). It is maintained that economic information obtainable in news articles is closely associated with social facts because opinions are articulated in forums, reviews or blogs.

Furthermore, numerous jobs related to marketing have become simpler to carry out. One example is marketplace research intended for business intelligence and competitive observation. New kinds of expression on online settings make it easier to gather important data that can be used to facilitate and discover transformations in market behaviour, to determine new technologies and markets where services and products are required, and also to detect threats.

Similarly, by means of opinion data, corporations are able to spot market segments and place their services and products in the most suitable areas; they are also able to equip themselves better to respond to their customers' needs. The investigations of the data flow on online social network-based platforms are able to identify dissimilarities among the

[7] RSS is a kind of formatting that can deliver regularly any changing in a Web content. This is a type of formatting which is capable of regularly providing any modification in Web content.

services offered by businesses and the opinions expressed by customers; they are also able to make comparisons between a company's abilities and those of its competitors. Finally, with an understanding of the huge amount of available information and of its linked opinions, businesses are given the opportunity to enhance their decision-making process through the discovery of novel facts and proposed solutions to their economic or technological problems.

According to Stavrianou and Chauchat (2008), the large amount of data available brings new potential to businesses. Conversely, a lack of social and financial data on marketplaces leads to incorrect or delayed decisions, and ultimately to significant economic losses. The result is the need for an automatic system offering the relevant essential information from a variety of sources magazines, newspapers, social network blogs, Internet sources, forums, and so on to facilitate the decision-making process (Conrad and Schilder, 2007; Pang and Lee, 2008; Stavrianou and Chauchat, 2008).

As stated by Pang and Lee (2008), the development of Web-based social networks and communication arrangements among their members is related to the building of motivating experiences, whose outcomes are together negative and positive. In online social networks, such as Twitter, people talk about the subjects that they would not deal with in their daily life, and they do this through their family and friends.

According to Balahur and Balahur (2009) as the rising volume of data permits key businesses and corporations and the general public to be more knowledgeable on 'what is happening' and/or 'what the world thinks about it', the quantity of data to be assessed cannot be handled manually. Therefore, dedicated systems have to be developed to collect and dig out the related opinion data and to offer appealing uses of this online content. The huge volume of data held in these resources necessitates the development of NLP systems in order to analyse this data systematically.

Therefore, recent years have seen the beginning and subsequent growth of a huge research initiative in the areas of NLP as it relates to practices normally categorised as sentiment analysis, opinion mining, review mining, subjectivity analysis or information extraction (Balahur and Balahur, 2009).

2.3.4 The role of opinion mining in the present business environment

In the past, organisations were dependent on emails, telephones, interviews and other slow methods to find out how customers felt about their products. According to Denecke (2008), with the advancement of micro-blogging, these organisations are now able to review their information through the use of online technology. Furthermore, customers can get in touch with the organisation through comments on social networks or in chat rooms; such a system is able to retrieve the required information from other Internet sources or through specific reviews on the Internet. The purpose of opinion mining and sentiment analysis is to come up with positive and negative opinions about a subject and to be able to match this information with information on other forms of social media, such as Twitter.

There are many messages on the Internet that offer assistance to Internet users, and many of these come from Twitter, Tumblr (www.tumblr.com), and Facebook (www.facebook.com) (Esuli and Sebastiani, 2005). Most of the users of such messages talk about their lives, share information that is vital to one another on a variety of subjects or matters, and talk about current problems in societies. According to Pak and Paroubek (2010) the micro-blogging websites are arranged in such a way that Internet users are able to move easily from one communication tool to others. With people having different opinions about world politics, wars, world religions, and so on, and others advertising their products and services, micro-blogging websites have become reliable sources where people can air such views and share their feelings (Zhang et al., 2007).

As Takamura et al. (2004) state, sentiment analysis is important not only to the company itself, but also to the customer. The information obtained from such texts is vital in shaping the quality of service provided by various organisations to the consumer. Through various languages, the opinions of the Internet users have been analysed to determine how they feel about a given matter. Sentiment analysis includes, for example, subjectivity detection, polarity classification, review summarisation, humour detection, emotion classification and sentiment transfer (Takamura et al., 2004).

2.3.5 Applications of sentiment analysis in social networks

According to Devitt and Ahmad (2007a) and Devitt and Ahmad (2007b), a number of studies have been carried out in the areas of sentiment analysis and opinion mining. These studies are intended to enhance diverse financial, social, political and psychological areas of daily human life. There have been numerous implementations of sentiment analysis and opinion mining applications to real-world situations; some of these are currently accessible online, such as Wefeelfine (www.wefeelfine.org) and Swotti (www.swotti.com), while others are still under investigation. In addition, other guidelines and developments in the area are emerging.

Websites like Swotti offer facilities like mining data, classifying information and summarising judgments. These online systems can be used for data advice, comparison or simply suggestions. A number of other similar systems directly connected to commerce are competitive market places for businesses that employ sentiment analysis and opinion mining of the Web; these systems allow businesses to obtain sincere, direct and impartial marketplace feedback regarding their services and products, together with the services and products in the market which are in competition.

Businesses and public figures make use of opinion mining to check their public image and public standing. Authors are able to profit from opinion mining by assessing

their literary reputation. It has been confirmed that variation in public estimation leads to changes in market stock prices intended for targeted business companies (Devitt and Ahmad, 2007a; Devitt and Ahmad, 2007b).

2.4 The role of blogs in building public opinion: mining blogs

Recently, there has been a rush towards posting and reading online blogs among all age groups. A lot of people use blogs as a chance to correspond with others and to share their views, practices, opinions and attitudes on a variety of subjects. For example, these blogs can be on specific products, current events, businesses and other people. Using this approach, a lot of data and information demonstrating people's personal sentiments on a variety of subjects has been added to the Web.

According to Vechtomova (2010) there is a great deal of biased information about numerous subjects on the Internet. In this situation, people want to utilise it in different situations, for instance, when selecting a service or a product, or when making an investment decision. Though discovering what others believe is not always simple, the common search engines (e.g. Google) can retrieve numerous pages containing simply factual details, such as shopping information and technical documentation (Vechtomova, 2010).

In mining online blogs, we aim to get more and more information regarding better decision making. Blog mining is one of the main tasks these days for better assessment and analysis of people's views on business products or services. However, this is not an easy or straightforward job. Lots of complexities and issues arise because of the unstructured nature of the data. Zhang et al. (2007) state that there are many techniques which make use of a lexicon of subjective words plus phrases gathered automatically and manually developed resources. According to the same study, one method used involves

the use of variance to weigh subjective words appearing near query terms in a given document.

2.5 The role of Twitter in building public opinion: mining Twitter

Micro-blogging nowadays has become an extremely well-known collaborative and communication tool for Web users. Billions of messages are produced each day, and an increasing number of websites offer services for micro-blogging. One of biggest micro-blogging sites is Twitter. The Twitter online platform is utilised by millions of people each day to state their opinion on a wide range of subjects; therefore, it is an important source of people's views and opinions.

Twitter holds a huge number of text posts, and this number is increasing on a daily basis. Twitter's audience varies from individuals to corporations; celebrities to politicians; and even presidents. As a result, it gathers text posts of clients from widely dissimilar social and interest groups. Twitter's viewers are users from a large number of nations; this aspect is making Twitter fertile ground for the gathering of opinions in diverse languages (Bifet et al., 2011).

2.6 Summary

In summary, as mentioned at the beginning of this chapter, the purpose of this chapter has been to review the importance of our chosen research domain, opinion mining and sentiment analysis in social media; to define the scope of the research; and to end with the final target of our study, namely, Arabic blogs and Twitter sentiment analysis.

To this effect, the chapter emphasises the importance and potential benefits of making use of the data in such social media websites (the benefits for e-commerce, shopping, entertainment, etc.). The chapter also points out how Web-based social media in particular with its vast number of users and huge amount of expressed opinion, offers

a rich supply of raw, unstructured data. If this data is to be of any use, then it needs to be structured and properly managed, so that databases in the form of KDD, for example, could be used to extract information through IR (for example, the frequency of certain words) and the eventual extracted data (for example, the opinions of customers about certain goods).

The chapter argues that such information and knowledge could be achieved through text and opinion mining techniques and tools (through the establishment of, for example, lexicons) that eventually categorise data into useful domains or patterns. This latter process would make sentiment analysis (one of our research aims) an achievable task. We talked about how this sentiment analysis could be measured (subjectivity measurement) and its strength be assessed. We talked about how such sentiment is achieved through language-based statistical interrogatory processes.

We have touched upon the fact that there are many available methods and tools available that are capable of achieving such aims; however, we made clear the deficiency and, sometimes, the non-existence of such methods or tools that target Arabic-based social media websites. We explained the difficulties encountered with the use of the Arabic language in such systems, compounded by transliteration issues (diglossic ones, for example). We discussed the need to make use of existing English-language-based methods and tools, and adapt them for our purposes.

We finally expounded on the availability of social media websites that people, especially Arabic speakers, use in their day-to-day social media communication, singling out blogs and Twitter (our research scope) as our target; a new approach towards amalgamating opinions between these two social media websites and the subsequent extraction of useful information and knowledge can be achieved.

CHAPTER 3: RESEARCH METHODOLOGY

Our research framework (RO_1) had four phases, as shown in Figure 1.2, which were analysed, implemented and evaluated. Each phase is described in detail in Chapters 6, 7, 8 and 9 respectively, while here in this chapter, we give a brief overview of our research framework. Figures 3.1, 3.2, 3.3 and 3.4 simplify the workflow of our data through these four phases to enable a clear understanding of our research.

3.1 Phase 1: corpus data processing

Each blog post was processed in order to extract multi-words terms (MWTs) using ArTerMine; this is illustrated in Figure 3.1. For more details, see Chapter 6.

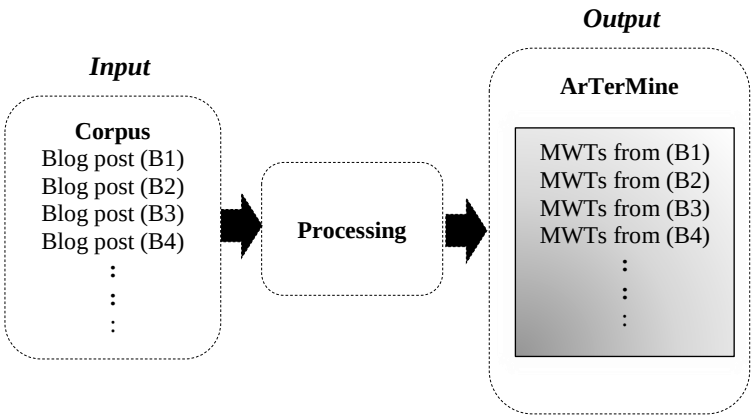


Figure 3.1. *Simplified overview of Phase 1: corpus data processing*

3.2 Phase 2: blog posts clustering and similarity

Each blog post was analysed using the implementation of Latent Dirichlet Allocation (LDA) topic modelling offerd by MALLET (http://mallet.cs.umass.edu); our hybrid

method includes three levels. As shown in Figure 3.2, we applied LDA topic modelling to cluster: (a) the raw text from a blog post to yield clusters of single words; (b) all extracted MWTs related to this blog post to yield clusters of MWTs; (c) both together to yield clusters including both single words and MWTs. Finally, as we sought only useful clusters, a similarity filter was applied; this would evaluate the entire collection of clusters to identify clusters with high similarity scores. In this thesis, a "cluster" means a group of words and/or multi-word phrases, extracted from a text such as a blog post; and "clustering" means extracting significant words and phrases from a text.For more details, see Chapter 7.

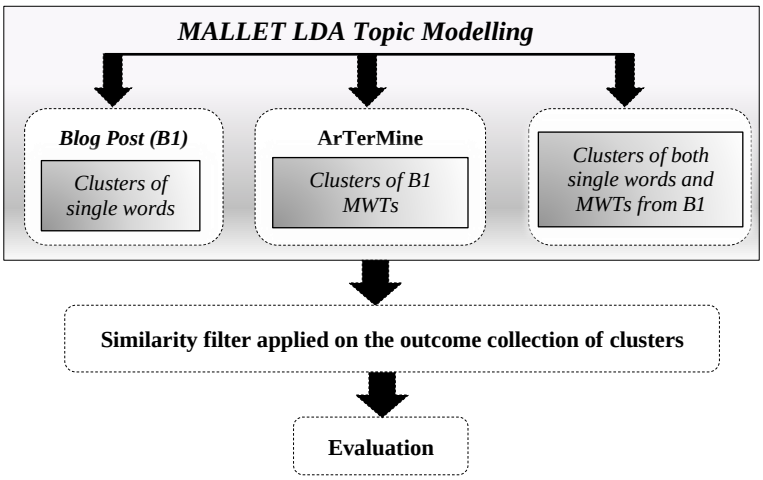


Figure 3.2. *Simplified overview of Phase 2: clustering of blog post words (B1) and MWTs of (B1)*

3.3 Phase 3: analysis of Twitter data

To collect related tweets for each blog post, we used terms that had been extracted from ArTerMine by using them as search terms in a keyword search using the Twitter API and collecting all related tweets retrieved. As the ArTerMine terms were in the transliterated format, we implemented an Arabic transliteration converter to change the text to a readable format. For more details, see Chapter 8.

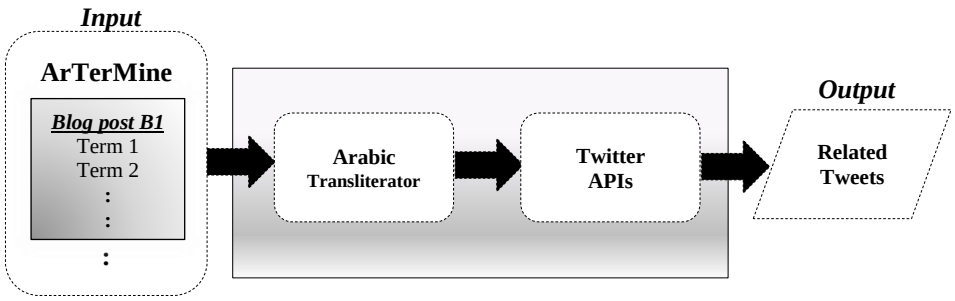


Figure 3.3. *Simplified overview of Phase 3: analysis of Twitter data*

3.4 Phase 4: linking, sentiment classification and ranking

As stated in our (RO3, RC1) and as shown in Figure 3.4, the useful clusters (see section 3.2) generated from each blog post and its related tweets were used as input; by this, we mean that our analysis pipeline in this phase applied per post separately. The ASWN and the TechTerms list were applied at the lexicon look-up stage to determine words related to sentiments and to technology. The collection of useful clusters and related tweets was applied at the next stage to classify sentiments in both. Then, the sentiments were ranked, and the final evaluation was carried out. For more detail, ranking of the degree of sentiments and the final evaluation, see Chapter 9.

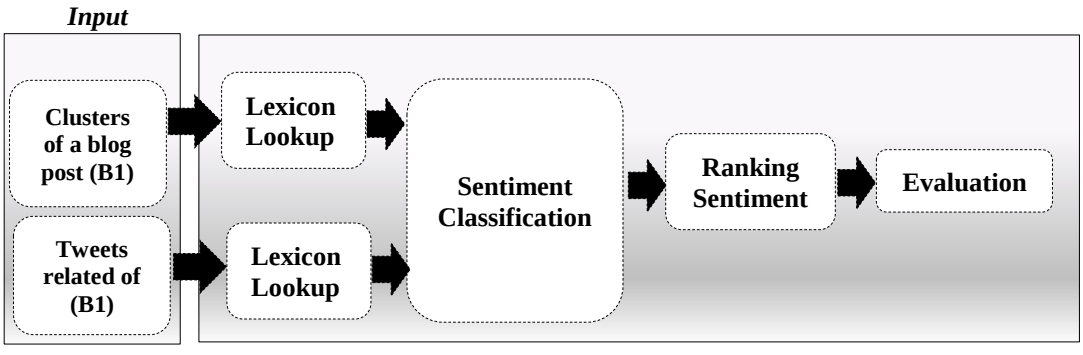


Figure 3.4. *Simplified overview of Phase 4: linking, sentiment classification and ranking*

3.5 Summary

The research framework was designed to help us focus on the most important stages required. In this chapter, we covered briefly the methodology for our research. An overview of what we have done was outlined.

Our research framework consisted of four distinct phases: firstly, corpus data processing; secondly, clustering of blog posts and similarity; thirdly, Twitter data analysis; and finally the fourth phase, the process of linking, classification of sentiments and ranking tweets based on sentiment degree, was carried out. In all these phases, we carried out the following: analysis, implementation and evaluation.

In Phase 1, the analysis was performed on the data that we have gathered from blogs, eventually extracting MWTs using ArTerMine. In Phase 2, we analysed each blog post using MALLET LDA topic modelling; introducing our hybrid method in the process. Raw text from blogs, the extracted MWTs, and a combination of the two together were processed, independently, to pick up useful clusters through a similarity filtering process. Finally, we evaluated the clustering outputs. In Phase 3, we collected and analysed the tweets; using the MWTs extracted from ArTerMine. We made a use of our Arabic transliteration converter for the transliterated terms. In Phase 4, the linking and sentiment classification processes were carried out. ASWN and TechTerms were used to determine the relevant words in our lexicon, after careful analysis.

CHAPTER 4: GENERATION OF THE RESEARCH CORPUS

In this chapter, we explain the development of our research corpus (RC2), covering its design, data characteristics, human evaluations, etc. But first, we provide a short summary of orthographic and morphological characteristics of written Arabic, the language of the corpus. Additionally provided are transliterations as well as forms of Arabic. The chapter will then discuss the design of our corpus, which was generated from technology blogs written in Arabic. The chapter will highlight the challenges and the difficulties faced through the stages of data preparation, corpus annotation, and corpus evaluation.

4.1 Characteristics of the Arabic language

4.1.1 A summary of the history

Lipiński (2001) describes the introduction of the phrase 'Semitic language' in 1971 by German historian A.L. Schloezer; it was the tongue of the biblical *Sem* (*She*), which had a past extending over 45 centuries. This family comprises a fraction of the Afro-Asiatic family, and its original written state was recorded during the third millennium BC. The Semitic languages were among the first languages to acquire a written form, with Akkadian writing commencing during the middle of the third millennium BC. Currently, Arabic is the most broadly used Semitic tongue, before Amharic, Hebrew and Tigrinya (Hetzron, 1997).

Versteegh (2001) outlines the key features of the Semitic language family; a Semitic language should comprise a root–sequence morphological structure, the presence of emphatic/glottalised consonants, and a verbal structure comprising a prefix and suffix conjugation. All of these elements must be present if a language is to be categorised as Semitic. Arabic is the mother language of over 317 million individuals in the Arab world.

The UN estimates that by 2020, over 395 million people will be resident in Arab nations (UN Development Programme, 2009). Furthermore, for over 1.5 billion Muslims globally, Arabic is the religious liturgical tongue. The UN has ranked it sixth in terms of significance, and it is one of the UN's official tongues. It was also ranked in fifth place by Weber (1997) in his piece on the ten most influential tongues; this ranking is based on the number of primary and secondary speakers, the economic influence of those nations employing the language, the number of key regions of human activity where the language is significant, the number and populations of nations using the language, and the socio-literary standing of the language. As it encompasses a time frame between 1980 and 1990, this research is comparatively old; however, Weber (1997) considered that his finding required no updating as the population of the world had increased comparatively.

As recorded by an official Internet observing agency[8], the number of users of the World Wide Web grew by 445 percent from 2000 to 2010. Remarkably, the Middle East, not including African nations, recorded the second-highest rate of increase in the number of Web users during the same time frame: nearly 1,825 percent; this is just below Africa, which comprises ten Arab nations, at nearly 2,800 percent. If we assume that the majority of these users speak Arabic, there will need to be a radical change in Web content for Arabic-speaking users; it will be necessary to address the lack of Arabic NLP resources and instruments. Of late, endeavours to augment Arabic content on the Web have been initiated, such as King Abdullah's endeavour, arranged by KACST,[9] which will enhance the use of the Internet by Arabic speakers.

[8] www.internetworldstats.com
[9] http://www.econtent.org.sa/Pages/Default.aspx

4.1.2 Orthography of the Arabic language

Arabic is written from right to left, like other members of the Semitic family (Tigrinya, Hebrew and Amharic). It is comprised of 28 letters in fundamental forms which encompass three extended vowels. Non-basic forms comprise letters produced by an integration of two letters. Furthermore, five key short vowels which do not make up part of the alphabet are included as diacritics and there are 13 brief vowel combinations overall. Vowels and short vowels correspond in the manner that a vowel comprises a double short vowel. These diacritics are employed mostly for the precise articulation of consonants, which alternatively assists in highlighting the precise translation. They are positioned over or underneath letters. This procedure is detailed as vocalisation and text may be completely, partially or never articulated subject to the written state (Buckwalter, 2004).

In Arabic, the letters are comprised of a cursive feature, signifying that a letter could be comprised of a varying shape subject to its position within a word: primary, medial, final or individual location. Furthermore, within the alphabet only six letters have two feasible states as only previous letters can link to them; these six letters cannot be linked with succeeding letters (Abdelali, 2004).

In Arabic writing, one key element is the absence of capitalisation, signifying that orthographic variations regarding case are not displayed. An additional aspect of Arabic is that there are fewer punctuation marks than in other languages, although of late these have been included. The Kashida [–] comprises a unique character used to lengthen a letter. For example, extending the letter [ح , H, h] as part of the word [محمد , mHmd, Mohammad] provides the new state [محمد]. Its application is for either deferring to the limitations of calligraphy or justification of text (Elyaakoubi and Lazrek, 2005). Arabic is the second most commonly employed script; it is employed in Sindhi, Farsi, Kurdish, Urdu and Pashto (Wagner et al., 1999).

4.1.3 Transliteration

Arabic characters need a means of portrayal by Latin characters within the digital age as a result of the absence of support for Arabic characters in the majority of computer software. The other cause is that readers who do not comprehend have a better comprehension when outlining the elements of the language. Transliteration is the one-to-one mapping from the source language to the intended script, which comprises part of Romanisation to translate spelling and not be perplexed by transcription to portray pronunciation.

Several schemes have been employed for translation within literature, e.g. Buckwalter (Buckwalter, 2004), LC (Banry, 1997) as well as ISO (Stone, 2001). The Buckwalter scheme as outlined by Figure 4.1 was selected by this research for its ease of use and because it does not entail any strange diacritics. Furthermore, the majority of the literature concerning Arabic NLP has applied this scheme, which offers more consistency for this research. Further particulars of this scheme are available on the Qamus[10] website, such as the Unicode characters.

[10] http://www.qamus.org/transliteration.htm

ء '	ذ *	ل l
آ |	ر r	م m
أ >	ز z	ن n
ؤ &	س s	ه h
إ <	ش $	و w
ئ }	ص S	ى Y
ا A	ض D	ي y
ب b	ط T	ً F
ة p	ظ Z	ٌ N
ت t	ع E	ٍ K
ث v	غ g	َ a
ج j	ـ _	ُ u
ح H	ف f	ِ i
خ x	ق q	ّ ~
د d	ك k	ْ o

Figure 4.1. *Buckwalter transliteration scheme*[11]

4.1.4 Morphology of the Arabic language

Arabic has a rich morphological structure, in which words are marked clearly for case, voice, gender, tense, number, definiteness and alternative morphological elements (Maamouri et al., 2006). As a highly inflected language, it relies heavily on prefixation, suffixation, and derivational and inflectional procedures. As well as affixation, it comprises clitic connection to stems. Following is a review of Arabic word classes and the manner in which they are produced (Kiraz, 2002).

[11] http://www.qamus.org/transliteration.htm

4.1.5 Forms of the Arabic language

A range of varieties of Arabic are employed in different contexts. All varieties originate from classical Arabic, the language of the Qur'an. Classical Arabic has good structuring and is completely vowelised although seldom used at present. The variety of Arabic currently employed for the majority of non-formal interaction activities is colloquial (regional or dialect) Arabic. Dialects have less inflection than the classical language; for instance, 'أنتم antm', a masculine plural personal pronoun, is employed to refer to both sexes in Jordan instead of 'أنتن antn' from classical Arabic. Although less commonly written, colloquial Arabic is becoming more common in poetry and on Web forums.

The currently employed official language is a type of diglossia, or the use of two varieties of the language (Farghaly and Shaalan, 2009). In the Arabic-speaking world, Modern Standard Arabic (MSA) is commonly used, integrating both colloquial and classical forms. In MSA, the key element is the complete or partial lack of diacritic indications which signify vowels. Abdelali (2004) details varieties of MSA and identifies lexical variations in MSA in ten nations as loan words and spelling variations among words.

At present, a critical aspect of Arabic writing comprises spelling errors in Arabic. Shaalan et al. (2003) explored the origins of common spelling errors, classifying them as errors of morphology, hearing and writing, among others. Within the written text of all tongues, spelling errors abound. Nonetheless, in Arabic, some errors have become common practice in writing. However, this practice is crucial when the error could lead to an entirely different analysis. For instance, if the word 'walker' was erroneously typed as 'walked' in English, the sense is not altered considerably and this may have no effect towards the processing at a certain level. The shapes of letters and sounds render Arabic more sensitive to errors.

4.2 The research corpus

4.2.1 Corpus design

A corpus is defined as follows:

> A collection of semiotic data that is stored electronically, devised in regards to a particular corpus criteria to represent, to be ultimately significant of a specific range of semiotic structure or alternative language (Butler, 2004).
>
> A collection of pieces of language text that is in electronic state, carefully picked in correspondence with external criteria to signify, as much as possible, a language or language range as an origin of data for linguistic study (Sinclair, 2005).

According to Sinclair (2005), the word 'pieces' is employed in the second definition due to the fact that a number of corpora are still using sampling techniques as opposed to gathering entire texts or a record of entire speech events. The initial purpose of the corpora is emphasised in order to set them apart from other language collections. There are three key principles that every corpus must take into account: representativeness, balance and the size of data (Biber, 1993; Sinclair, 2005).

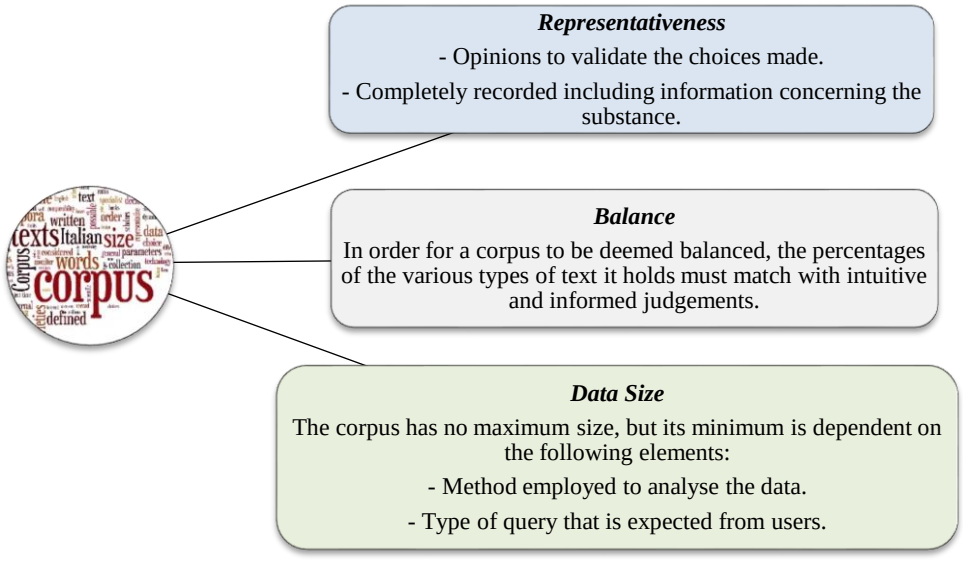


Figure 4. 2. *Corpus design principals based on (Butler, 2004)*

Considering these principles, we designed our corpus in such a way as to fulfil the purpose of this research. It is important to mention that, the balance criterion was not relevant and was not an issue in our research corpus because it refers to the variety of text types in the corpus whereas we built our corpus using one type of text (formal Arabic blogs websites). Regarding the representativeness criterion, as shown in Table 4.1, we considered Arabic technology blogs in which any blog post should represent and include the following types of data, in order to be used for our corpus (this shows the manual work required to set up our corpus):

1 Any type of the technology terms that represent: *English companies, English companies transliterated to Arabic, English technology, English technology transliterated to Arabic, Arabic companies or/and Arabic technology*.

2 Positive or negative opinions about the technology terms represented in these blog posts.

3 Factual information about these technology terms which do not express any sentiments.

Additionally, based on these types of data, Table 4.1 also illustrates the percentage of each type of data in our corpus (the data size criterion).

Table 4.1. *Research corpus design*

Types of Data		**%**
Technology Terms	English companies	30%
	English companies transliterated to Arabic	
	English technology	
	English technology transliterated to Arabic	
	Arabic companies	
	Arabic technology	
Sentiments	Positive or negative opinions about technology	47%
Facts (Neutrals)	Information about technology which does not express any sentiments	23%

4.2.2 Corpus preparation and challenges

Our corpus is based on formal MSA Arabic and includes 41 different Arabic blog posts (the total number of sentences is 2,350 sentences). As stated in Alhazmi and McNaught (2013) five well-known Arabic technology blogs were used for our corpus: *Technology-world*[12], *Unlimited-technology*[13], *Teedoz*[14], *Apple-world*[15] and *Notebook-technology*[16]. There were two reasons for choosing Arabic technology blogs written in formal Arabic:

1 Generating a novel corpus for Arabic sentiment analysis which covers a new domain (Arabic technology blogs) in comparison to the existing corpora for Arabic sentiment analysis which focused on e.g.:

- *Movie reviews:* as used by Rushdi-Saleh, et al. (2011) in their corpus for opinion mining for Arabic (OCA).
- *Datasets from Wikipedia Talk Pages, Penn Arabic TreeBank and Web forum:* as used by Abdul-Mageed and Diab (2012a) who described a multi-genre corpus for Arabic sentiment analysis (AWATIF).
- *Dataset from newswire*: as described in Abdul-Mageed, et al. (2011).

2 The corpus provided us with several challenges. Table 4.2 shows these challenges and our proposed solutions, taking into consideration the Buckwalter transliteration scheme, shown in Figure 4.1, to represent our examples.

[12] http://www.tech-wd.com/wd/
[13] http://www.unlimit-tech.com/blog/
[14] http://www.teedoz.com/
[15] http://www.apple-wd.com/
[16] www.tech-k.com/

Table 4.2. *Corpus challenges and proposed solutions (Alhazmi and McNaught, 2013)*

Challenges	Proposed Solutions	Examples
Foreign language – Roman alphabet used	Building an open list including technology, company and software names.	Google – Facebook – Samsung – Galaxy S3 – iPhone, etc.
Transliterations	Attaching all transliterations to our list, and if there is more than one transliteration for a word, all of them needed to be indexed and included.	Google: - qwqwl قوقل - jwjwl جوجل - gwgwl غوغل Facebook: - fAysbwk فايسبوك - fysbwk فيسبوك
Misspelling	Data was checked by experts in the Arabic language.	Quality: - jwdh جوده - jwdp جودة

4.2.3 Human annotation of corpora: guidelines

There are three key procedures in developing a gold standard of annotation (Hirschman and Mani, 2003):

1 Identifying what must be annotated and ways in which to annotate it by employing annotation regulations.

2 Identifying instruments to assist with the annotation procedure, particularly in the contrasting of the scores.

3 Being able to validate the results through calculations employing inter-annotator agreement.

Inter-annotator agreement is employed to produce statistical values signifying the precision of the annotations, as all aspirations of automating a procedure like this are lost if people have challenges in establishing what to annotate. For that reason, the Kappa[17]

[17] Kappa is a statistical tool to record inter-annotator agreement of qualitative items (Carletta, 1996; Fleiss et al., 2013).

coefficient or an absolute rate of agreement can easily be employed in order to stipulate the statistical value (Kim and Tsujii, 2006).

Our corpus is viewed as a gold standard for this research, ready to be analysed and annotated. A small team of three annotators was used to identify and annotate our datasets for technology blog posts; the annotation guidelines used are shown in Table 4.3, and Figure 4.3 shows examples of the annotated corpus. Furthermore, Table 4.4 illustrates the annotation guidelines with examples.

Table 4.3. *Annotation guidelines for technology blog posts*

Annotation Guidelines		
TechTerms (Technology Terms)	***Facts (Neutrals)***	***Sentiments***
Company names – written in Roman alphabet.	Facts about companies.	Positive or negative opinions about companies.
Company names – transliterated into Arabic.		
Company names – written in Arabic.		
Technologies, products or systems – written in Roman alphabet.	Facts about technologies, products or systems.	Positive or negative opinions about technologies, products or systems.
Technologies, products or systems – transliterated into Arabic.		
Technologies, products or systems – written in Arabic.		

```
<?xml version="1.0" encoding="utf-8"?>
<Arcaptcha>
<English_technology_name>Arcaptcha</English_technology_name>
<Arabic_technology_name>الكابتشا</Arabic_technology_name>
<Positive_opinions>
<Positive_opinion>مشروع عربي هو الأول من نوعه في عالمنا</Positive_opinion>
<Positive_opinion>شخصيا أدعم هذه الفكرة و أؤيدها بشده</Positive_opinion>
</Positive_opinions>
<Negative_opinion> ولكن العديد منا أصابه الملل من الحروف اللاتينية التي كنا نحن كعرب مجبرين على استخدامها كل
مرة</Negative_opinion>
<Facts>
<Fact>تطوير لتقنية كاباتشا من خلال دعمها للغة العربيه</Fact>
<Fact>سكتبة برمجية لمطوري المواقع والانظمة المتصلة بالشبكة للتصدي للطلبات الوهمية والاوتوماتيكية</Fact>
<Fact>اختبار يميز بين إجابة المستخدم الأدمي وبرامج الحاسوب</Fact>
<Fact>فكرة المشروع الجديد بهدف تكثيف المحتوى العربي و تسهيل إيصاله لجميع المستخدمين العرب</Fact>
</Facts>
</Arcaptcha>
```

```
<?xml version="1.0" encoding="utf-8"?>
<CSS>
<English_technology_name>CSS</English_technology_name>
<Arabic_technology_name>سمة</Arabic_technology_name>
<Positive_opinion> طريقة جميلة لتسهيل تطوير صفحات ومواقع الإنترنت على هؤلاء الذين يعانون من ضعف في اللغة
الإنجليزية</Positive_opinion>
<Negative_opinion>
</Negative_opinion>
<Facts>
<Fact> عبارة عن النسخة العربية او تعريب لغة CSS </Fact>
<Fact> تمكنك من كتابة ملفات التنسيق بلغة عربية </Fact>
<Fact>مستخدمة ملف جافا سكربت الذي ستقوم باستخدامه CSS ستتولى سمة تحويل نفسها إلى تنسيق</Fact>
</Facts>
</CSS>
```

Figure 4.3. *Examples of the corpus*

There was a high rate of inter-annotator agreement in determining the TechTerms task, especially for marking company names (written in English, transliterated into Arabic or written in Arabic). However, disagreement occurred in distinguishing between facts and sentiments (Alhazmi and McNaught, 2013). An example of each task will be reviewed here, while the discussion of the results takes place in section 4.3.

One of the challenges presented by corpus research is writers using a foreign language—in this case, English—to represent companies, technologies, products and so on. A further challenge is presented by writers using transliteration from English to Arabic. Table 4.4 shows examples of each of the tasks assigned to the annotators as highlighted data in the table, which is shown above in Table 4.3.

Table 4.4. *Annotation tasks with examples (Alhazmi and McNaught, 2013)*

Assigned tasks	Examples in Arabic with Buckwalter transliterations
Task1 (TechTerms): an English company name in Roman alphabet	اتمام صفقة استحواذ جوجل على Motorola Mobility tmAm Sfqp AstHwA* jwjl ElY Motorola Mobility
Task1 (TechTerms): an English company (e.g. **Google**) transliterated into Arabic	اتمام صفقة استحواذ جوجل على Motorola Mobility tmAm Sfqp AstHwA* jwjl ElY Motorola Mobility
Task1 (TechTerms): an Arabic company	شركة حسوب تكشف النقاب عن إعلانات الفيديو وتعزز منصتها الإعلانية بمتجر إعلانات $rkp Hswb tk$f AlnqAb En <ElAnAt Alfydyw wtEzz mnSthA Al<ElAnyp bmtjr <ElAnAt
Task1 (TechTerms): an English technology	ياهو تطلق متصفحها الجديد Axis yAhw tTlq mtSfHhA Aljdyd Axis
Task1 (TechTerms): an English technology (*sky drive*) transliterated into Arabic	مايكروسوفت تقول أن سكاي درايف هي الأفضل على الإطلاق mAykrwswft tqwl >n skAy drAyf hy Al>fDl ElY Al<TlAq
Task1 (TechTerms): an Arabic technology	مغردون :موقع لنشر تغريداتك على تويتر وتسويقها mgrdwn mwqE ln$r tgrydAtk ElY twytr wtswyqhA
Task2 (Facts): about an Arabic technology	مغردون :موقع لنشر تغريداتك على تويتر وتسويقها mgrdwn mwqE ln$r tgrydAtk ElY twytr wtswyqhA
Task3 (Sentiments): a positive and negative sentiment	خطوة كبيرة من الفيس بوك بتوفير هذا المتجر والذي سينجح بكل سهولة xTwp kbyrp mn Alfys bwk btwfyr h*A Almtjr wAl*y synjH bkl shwlp أرى أن مستقبل نظام الاندرويد بدأ يصبح مخيفاً وغير آمناً >rY >n mstqbl nZAm AlAndrwyd bd> ySbH mxyfAF wgyr |mnAF

4.2.4 Evaluation of the human corpus annotation

The Kappa statistic is used to evaluate the level of agreement for corpus annotation by humans. The results obtained are reported based on the interpretation of Kappa, as shown in Figure 4.5.

	Poor	Slight	Fair	Moderate	Substantial	Almost perfect
Kappa	0.0	.20	.40	.60	.80	1.0

Kappa	Agreement
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21– 0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

Table 4.5. *Interpretation of Kappa (Viera and Garrett, 2005)*

In Kappa, the inter-annotator agreement of qualitative items is recorded by a statistical tool (Carletta, 1996; Fleiss et al., 2013). The qualitative items are a pair of annotators, each of which categorises *M* items into *S* mutually exclusive classes; this was discussed by Smeeton (1985). The formula used to calculate κ is:

$$\boldsymbol{\kappa} = [\boldsymbol{P}(\boldsymbol{a}) - \boldsymbol{P}(\boldsymbol{e})] / [\mathbf{1} - \boldsymbol{P}(\boldsymbol{e})] \quad \textbf{(4. 1)}$$

Where:

- *P (a)* = the probability of agreement,

  = the total number of agreements / *the total number of annotations*.
- *P (e)* = represents the probability of chance agreement, i.e. of each class being observed at random.

K varies between −1 and 1. If the annotators agree fully, $\boldsymbol{\kappa} = \boldsymbol{1}$. If the annotators only agree to the extent to be expected from chance (P (*e*)), κ = 0.

Hence, each pair of annotators for all of the tasks required (TechTerms, facts and sentiments) was calculated by Kappa. The average of the Kappa results is shown in Table 4.6 and Figure 4.4.

Table 4.6. *The average agreement (Kappa statistic)*

Tasks	IAA 1&2	IAA 1&3	IAA 2&3	Kappa Average
TechTerms	0.809	0.914	0.881	0.868
Facts(Neutrals)	0.939	0.722	0.705	0.789
Sentiment	0.716	0.749	0.863	0.776

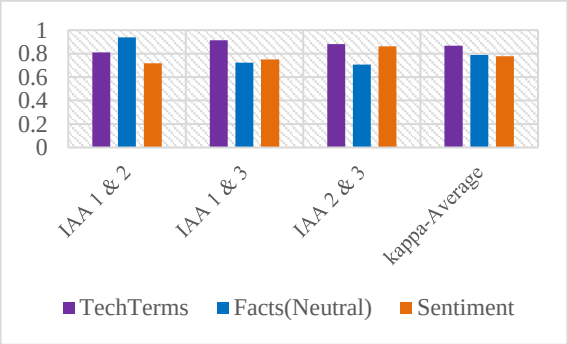


Figure 4 4. *Human evaluation of the corpus*

4.3 Discussion of the results

The inter-annotator agreement generally showed substantial agreement in the Facts and Sentiments tasks and an almost perfect agreement in the TechTerms task, with a Kappa score of 0.868; this met our expectation. However, we expected the results for the Facts and Sentiments tasks to be in the same range of Kappa scores, whereas the results obtained were 0.789 and 0.776 for Facts and Sentiments respectively. This is considered in the range of results shown in various other studies, which ranged from 0.7 to 0.8 (Abbasi et al., 2008a; Bifet and Frank, 2010; Kim and Hovy, 2004).

Nonetheless, as we sought high sentiments scores, because it is the core of this research, we thought of our *RQ2*. With the absence or lack of resources and tools for Arabic in the domain of sentiment analysis, we are trying to contribute to the building of the ASWN database in order to achieve improvements in sentiment scores.

4.4 Summary

As the object of this chapter was to present how we went about producing our corpus, it was prudent to give an initial brief relevant account about Arabic, its history, the issues pertaining to its orthographic and morphological features, especially in the written form of the language, that present researchers with problems, not faced by many other languages.

Since we were aiming at generating a new corpus for the domain of Arabic sentiment analysis, our study scope was technology blogs. We designed and constructed a corpus in this domain comprising 2,350 sentences. This was after taking into account the three main issues for the creation of any corpus - namely, representativeness, balance and size of data - as well as being dependent on the data analysis method employed and the type of expected query. This was done so that to take care of all those data which comprised of (*English companies names, as well as their transliterations to Arabic; English technology terms, as well as their Arabic transliterations; Arabic companies names; Arabic technology terms; polarity of opinions shown about technology; and finally neutral information, i.e. without any sentiment*).

The first step towards building the corpus was to build the human annotation corpora, however, there were three steps that need to be performed in order to achieve the Gold Standard annotation. They were: employ annotation guidelines; identify instruments for contrasting scores; and finally, validate through inter-annotation for statistical production purposes. The Kappa statistic (the purpose of which is to negate any agreement due to chance) and the absolute rate of agreement methods were employed, for the final validation process. As it turned out, we achieved a high degree of agreement (almost perfect) of inter-annotators in determining the TechTerms (the open list of technology terms, that we were compiled), but a lesser degree, although still substantial, of agreement

in distinguishing facts from sentiments. This must be taken in prospective, as it was in the range of results shown by other studies.

In order to create an automated text mining pipeline able to classify Arabic text units for sentiment in a comparable way, some language resources will be necessary, as we discuss in the next chapter.

CHAPTER 5: TOOLS AND LEXICON-BASED APPROACHES

After we generated our corpus, the data was ready to be analysed in various ways, both as a collection, and as individual texts. In this case, text mining tools and techniques were used. In this chapter, we mention briefly text mining levels; then, we explain in depth our contributions (RC_3, RC_4 and RC_5) to bringing about improvements in lexicons and tools for Arabic sentiment analysis. Given the lack of a sentiment lexicon for Arabic, we had to build the ASWN for use in our research. Furthermore, the ArTerMine tool was used for extracting Arabic terms; this tool was an extension of the TerMine tool, the English version. Finally, to meet the need for a small lexicon that contained technology terms, we made an open list.

5.1 Text mining levels

There are various levels on which any information from an unstructured text (a free text) can be extracted by mining; we started with information provided by our annotation guidelines and then tried to find this information within the texts; later, the research hypotheses were tested based on the information obtained. These mining levels, as shown in Figure 5.1, are (IR) Information Retrieval, (ER) Entity Recognition, (IE) Information Extraction, Text Mining and Integration (Jensen et al., 2006).

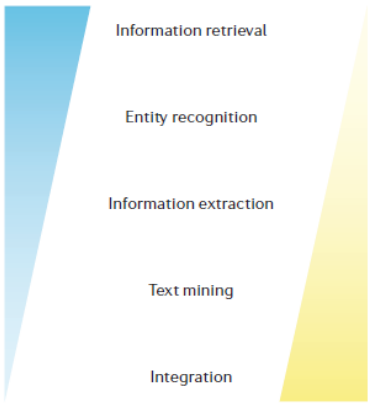


Figure 5.1. *Mining Levels (Jensen et al., 2006)*

For the purpose of our research, we made use of Information Retrieval (IR), Entity Recognition (ER) and Information Extraction (IE).

- ***Information Retrieval (IR)***

Information retrieval (IR) involves collecting all of the text that is to be mined. The text can be acquired from a variety of information sources, such as journal articles, website entries, social network, blog posts and more. An IR search operation can be regarded as a filter that returns a subset of the texts in a collection that match a given set of search terms. Furthermore, IR sometimes involves specifying particular sections of the texts that the researcher has an interest in. This could go down to the level of detail of searching for particular passages or perhaps even phrases (Jensen et al., 2006). In our research, we aimed to retrieve tweets from Twitter, which were relevant and which discussed the same issues as the blog posts. Thus, the process of gathering relevant tweets made use of IR in this case. The process of filtering down may be difficult and does not always achieve the intended aims. The researcher must make a choice between the slow process of manual filtering and the quicker, but potentially less accurate, approach of automated filtering. Also, it has generally been accepted that the mining must work from all old tweets, as a

limited dataset might miss relevant posts (Altman et al., 2008; Cohen et al., 2010; Seeber, 2008).

Luo et al. (2012a) tried to retrieve Twitter posts that met two key requirements:

1. They needed to be relevant to the study, blog posts in our case.
2. They needed to make specific reference to the area of study, regardless of what they said about it.

As far as Luo et al. (2012b) were aware, theirs was the first study to create a ranking model for the purpose of facilitating comment retrieval from tweets. They collated various opinions from Twitter regarding a particular field of interest. Their findings suggested that opinion retrieval works better when hyperlinks, references, and tweeter information (e.g. number of tweets/followers/etc.) and the point of view expressed in the post are all considered. In addition, it was suggested that the collected data is used to produce automated subjective tweets and objective tweets. This produced similar results to those of the manual approach (Luo et al., 2012a; Luo et al., 2012b).

- ***Entity Recognition (ER)***

After the corpus is ready, recognising the variety of named entities (NER) is the next step. Named entities are substrings consisting of words which are mapped onto specified groupings (Zweigenbaum et al., 2007). NER often uses trained components to locate and classify NEs that are not in the current dictionary, and these classifiers often use orthographic features such as capitalization, which is not available in Arabic.

In our research, these NEs can be technology, product or company names in both the Arabic and English languages. The most basic NER involves dictionary-based NER, for example, linking ‘TechTerms’ words from the lexicon to the text.

- ***Information Extraction (IE)***

Information Extraction (IE) comprises an element of a greater issue which addresses the challenge of the creation of automatic techniques for management of text, beyond its conveyance, storage and display (Freitag, 2000; Jensen et al., 2006).

The field of IR has established automatic techniques, characteristically of a statistical nature, to index and categorise great collections of documents. An alternative approach is known as *natural language processing* (NLP), which has dealt with the challenge of modelling the processing of human language with substantial success when the size of the task is considered (Freitag, 2000; Hirschman and Mani, 2003; Verma et al., 2016).

IE addresses tasks in between NLP and IR. Regarding input, IE supposes the availability of a group of documents within which every document observes a template, i.e. details one or several occurrences or items in a way that is comparable to other documents although varying in terms of details (Verma et al., 2016).

For instance, we can consider a collection of newswire articles regarding Latin American terrorism, where every article is assumed to have one or more terrorist acts as its foundation. We define for every provided IE task a template (Linguistic Data Consortium, 2005a; Linguistic Data Consortium, 2005b), comprising a (or a group of) case frame(s) to contain the information withheld in one document. Using the example of terrorism, a template would comprise slots relating to the victim, the perpetrator, the weapon used in the event and the date on which the occurrence took place. For this challenge, an IE structure is necessary to ‘comprehend’ an attack article sufficiently to discover information relating to the slots within this template (Freitag, 2000; Jensen et al., 2006).

5.2 ASWN database

Opinion mining involves estimating whether a text is subjective or objective and in the case of being subjective, whether it is 'positive' or 'negative' (Esuli and Sebastiani, 2006; Liu, 2011; Pang and Lee, 2004; Wilson et al., 2004). Opinion mining involves establishing a text's polarity, as well as the intensity of polarity, if there is any. There can be strong, mild or weak positivity or negativity of text, and such variations may be greatly applicable to the deductions that may be made from the appraisal. Given the innately subjective character of individual viewpoints, appraising the standard of the outcomes produced from any instruments causes specific challenges (Baccianella et al., 2010; Esuli and Sebastiani, 2006).

The creation of the SentiWordNet (SWN) 1.0 and 3.0, an English application that is accessible by the public and employed in opinion mining and sentiment categorisation, as well as the efficiency of the English versions, was detailed and contrasted in Alhazmi et al. (2013). The SWN is a developing resource which maps to subsequent WordNet (WN) versions (Miller et al., 1990).

Application of the SWN may be done using various languages, although as it maps towards WordNet, it necessitates a suitable WN version for the particular language employed. For the instance of Arabic, it necessitates the creation of an Arabic version of WN 3.0 which is subsequently mapped towards the English WN 3.0. If this is not available, the construction of sentiment evaluation for Arabic texts will lag in an area that has considerable potential for text mining. With a basis in this detail, we collaborated on constructing the ASWN in relation to the most recent English SWN 3.0, with consideration for upgrading the Arabic WN 2.0 version to WN 3.0 (Alhazmi et al., 2013).

5.2.1 Background: related work

In absence of a previously available SentiWordNet for the Arabic language, a number of studies concerning sentiment analysis and opinion mining in Arabic have employed lists/catalogues of sentiment words to address their research requirements (Al-Subaihin et al., 2011; Farra et al., 2010; Rushdi-Saleh et al., 2011). Recently, there have been efforts towards building Arabic sentiment lexicons; these efforts can, for the most part, be divided into two classes: (1) making a connection between an Arabic and an English lexicon; and (2) implementing semi-supervised or supervised learning methods for Arabic resources.

SANA 'Arabic Subjectivity and Sentiment Analysis', an Arabic lexicon for sentiment and subjectivity, was put forward by Abdul-Mageed and Diab (2014). Previously available lexicons are integrated into this lexicon, which concerns mechanical machine interpretation, gloss matching and manual annotation over a number of resources including SANA (Graff et al., 2009) and THARWA (Diab et al., 2014). Around 225,000 entries are included within SANA, with several of these being replications, not diacritised and inflected, rendering the resource noisy and less practical. Furthermore, the standard of the resource is influenced by the fact that automatic interpretation does not employ the part of speech (POS) data.

Additional work observing translation techniques includes a study by El-Halees (2011), in which SentiStrength (Thelwall et al., 2010) was interpreted by using a dictionary in addition to manual correction. An additional example is SIFAAT (Abdul-Mageed and Diab, 2012b), a prior version of SANA although with more dependability regarding translation. A further lexicon was constructed by Elarnaoty et al. (2012), who manually interpreted the MPQA lexicon (Wilson et al., 2005). These early studies are generally characterised by a lack of sufficient entries and quality.

A sentiment Arabic lexical Semantic Database (SentiRDI) was developed by Mobarz et al. (2011) and Abdelrahman et al. (2014) through the use of a dictionary-based technique. The database is comprised of numerous inflected forms, i.e. it is not lemma-based. However, the authors reported unsatisfactory quality and intend to try different options.

Elhawary and Elfeky (2010) developed another Arabic sentiment lexicon. This lexicon was constructed employing a similarity graph, in which the edges are comprised of similarity scores. A key disadvantage is the reduced coverage of the lexicon. Furthermore, extension of the graph necessitates a considerable corpus with semantic annotations and polarity, which increases the sparsity.

Recently, SLSA 'A Sentiment Lexicon for Standard Arabic' developed by Eskander and Rambow (2015), has appeared and is very similar to our ASWN. The SLSA has the highest reported coverage (34,821 words). The building of the SLSA was founded on the connection of the AraMorph[18] lexicon with SentiWordNet, in addition to several heuristics and strong back-off. A comparative improvement of 37.7 percent was demonstrated by the SLSA over a contemporary lexicon when assessed for precision. Additionally, it outperformed it by a definite 3.5 percent of F-measure when assessed for a sentiment evaluation.

Table 5.1 summarises the related work mentioned in this section based on three issues: (1) SWN-Based: if any of these lexicons was built on or linked to any version of SentiWordNet databases, (2) the coverage of these lexicons, (3) their availability to the public. By comparing these lexicons to our lexicon the ASWN, it can be seen that the main issue behind building the ASWN was the unavailability of these lexicons except the SLSA (Eskander and Rambow, 2015). Furthermore, as we had already conducted our experiments using our ASWN, which satisfied the needs for a sentiment lexicon for our

[18] AraMorph is an Arabic morphological analyser and POS tagger. It is available at https://sourceforge.net/projects/aramorph/

research, it was too late for the SLSA to be included in our research as it had only recently been made available. However, further discussion about the SLSA will take place in Chapter 10.

Table 5. 1. *Sentiment Lexicons for Arabic*

Resources	***SWN-Based***	***Coverage***	***Availability***
SANA (Abdul-Mageed and Diab, 2014)	No	225,00	No
SentiStrength (Thelwall et al., 2010)	No	2310	No
SIFAAT (Abdul-Mageed and Diab, 2012b)	No	3,325	No
A lexicon by Elarnaoty et al. (2012)	No	40,000	No
SentiRDI (Mobarz et al., 2011) and (Abdelrahman et al., 2014)	No	Not given	No
An Arabic sentiment lexicon by Elhawary and Elfeky (2010)	No	Not given	No
SLSA (Eskander and Rambow, 2015)	Yes	34,821	Yes

5.2.2 Database implementation

Arabic comprises a broadly employed language with social-political as well as economic significance, thus it is normal for there to be considerable interest in tools and resources that facilitate extraction and appraisal of sentiment in Arabic texts. Much development work in the field has concentrated on creating resources that are equally available for English and Arabic. In our case, through the inclusion of sentiment information in the Arabic WN to produce the ASWN, the set of WN-based resources has been enhanced to the advantage of Arabic NLP researchers. This has the possibility of facilitating the direct development of upcoming opinion mining tools for Arabic texts.

For the implementation of ASWN, the main initial stipulation is a version of WN that it may operate on. An Arabic version of WN 2.0 was available (Black et al., 2006; Elkateb et al., 2006), *Arabic WN 2.0*[19], although there was no Arabic version of WN 3.0.

Thus, a database was constructed for ASWN, with consideration of all the levels of groundwork (Alzhami et al., 2013) as in the following:

[19] http://www.globalwordnet.org/AWN/

- *Level 1*: Upgrading of the Arabic WN 2.0 database version to 3.0 was done automatically through mapping (*see the mapping step below*) to the most recent English WN 3.0 database.
- *Level 2*: Each of the fields within the new database, Arabic WN 3.0, was checked manually and then all these fields within the Arabic WN 3.0 were compared with the English SWN 3.0. Only the fields within the Arabic WN 3.0 that both articulated sentiments and matched a word within the English SWN 3.0 were maintained to provide us with the Arabic version of SWN, and the rest were erased.

By means of this mapping, an Arabic SWN database was produced. The complete number of words available within the ASWN comprises around 10,500, encompassing nouns, verbs, adjectives and adverbs.

Two different methods were considered for the construction of the ASWN: (1) employing the database constructed as a multilingual system for implementation in English as well as Arabic settings; and (2) requiring the translation of all synsets[20] within the English version into the Arabic version, which latter was the method we employed.

The mapping step

The process to establish the mapping is described fully in Daudé et al. (2000). The *Relaxation Labeling algorithm*[21] is described by Daudé et al. (2000) as:

1 A set of variables (representing words, synsets, etc.), are dealt with in this algorithm.

2 Each variable can take one of a number of diverse labels (POS tags, senses, etc.).

3 In addition, there is a set of constraints which set out the compatibility or incompatibility of a mixture of pairs variable-label.

[20] Synsets: are collections of 'senses' for a word that clarify the particular sense in which it can be used. The words that clarify each sense are called the 'gloss'. These comprise compilations of 'senses' for a word, which elucidate the specific sense within which it may be employed.

[21] A brief description of the algorithm is mentioned here, for more details see Daudé et al. (2000).

4 The algorithm's aim is to discover for each possible label, and for each variable, a weight assignment, in order for (a) the labels of the same variable's weights to add up to one, and (b) the weight assignment to satisfy the set of constraints as much as possible.

5 In summation, the algorithm carries out constraint satisfaction in order to decipher a consistent labelling. The following steps are taken:

(1) Begin with a weight assignment that is random.

(2) Calculate the label of each variable's support value. Support is calculated in relation to the constraint set and the current weights for labels of the context variables.

(3) Increase the weight of the labels that are more compatible with the context (larger support) and reduce the less compatible ones (smaller support). The weight change must be in proportion to the received support from the context.

(4) In the case where a stopping/convergence criterion is met, stop. In any other case proceed to step 2. Daudé et al. (2000) made use of the standard of stopping when no more changes take place, even though it is possible to bring an end to relaxation processes (Eklundh and Rosenfeld, 1978; Richards et al., 1981) using more sophisticated heuristic techniques.

(5) The algorithm's cost is relative to the product of the number of constraints multiplied by the number of variables.

Conversion of Arabic WordNet from correspondence with version 2.0 to version 3.0 of Princeton WordNet

The later version of Princeton WordNet is expanded compared with the earlier version, so a resource mapped to the earlier can be mapped to the later WordNet without loss of information. For purposes of data conversion, it is most convenient to start from an abstract relational model of the data. In this, the base table (created from file *wn_s* in the

Prolog distribution of WordNet) represents the relation between words and word senses. An extract is shown in Table 5.2.

Table 5. 2. *Extract from word-sense table in WordNet in relational format*

synsetid	*wnum*	*word*	*sstype*	*snum*
100004475	1	organism	N	1
100004475	2	being	N	2
113954253	1	being	N	1
108436036	1	organism	N	2

Considering the monolingual Princeton WordNet, in the relational (or Prolog) representation, it is convenient to use the synsetid as the unique identifier of a sense, so that 100004475 represents the synonym set comprising sense 1 of 'organism' and sense 2 of 'being'. Tables expressing sense relations, such as homonymy and holonymy, then comprise sets of pairs of synsetids. When we come to mapping wordnets for other languages, we may encounter word senses that have no equivalent in English, or at least, which have not been recorded in the Princeton WordNet. Accordingly, instead of using the arbitrary synsetid, derived from the relative position of the data about the synset in a physical file layout, the Arabic WordNet project used an alternative candidate key as the identifier of a synset, <word,wnum>, that is, the first listed word in the synset and the sense number of that word which participates in the synset.

The mappings between Arabic synsets and corresponding English (WordNet 2.0) ones are represented in this notation, in the *link* table of the Arabic WordNet database, as follows:

Table 5. 3. *Link table from AWN showing Arabic to English synset translations*

Relation	*Arabic synset ID*	*English WN 2.0 synset ID*
equivalent	$A}ik_Aljilod_n1AR	echinoderm_n1EN
equivalent	$a>n_n1AR	concern_n1EN
equivalent	$a>n_n2AR	thing_n7EN
equivalent	$aAEa_v1AR	break_v46EN

The SentiWordNet 3.0 data is issued in a plain text file format. Table 5.4 shows the first line of data beneath column headings (POS='a' refers to adjective part of speech).

Table 5. 4. *One-line data extract from SentiWordNet 3.0*

# POS	*ID*	*PosScore*	*NegScore*	*SynsetTerms*	*Gloss*
a	00001740	0.125	0	able#1	(usually followed by `to') having the necessary means or skill or know-how or authority to do something; "able to swim"; "she was able to program her computer"; "we were at last able to buy a car"; "able to get a grant for the project".

The synset ID in Table 5.2 precedes the 8 digit ID seen in SentiWordNet 3.0 (Table 5.4) with a digit corresponding to the part of speech: in the case of an adjective, for example, the prefix digit would be '3'. To establish a correspondence between an AWN synset ID and a SentiWordNet 3.0 synset ID, we need to transitively map from AWN synset ID, through the AWN *link* table's English WN 2.0 synset ID, and from thence to the 8-digit offset-based synset identifier in WN 2.0, for which a mapping to the WN 3.0 identifier is available from Daudé et al. (2000). The correspondence between the AWN-style synset ID and the 8-digit identifier is available from Table 5.2, where the AWN symbol is formed by concatenating *word*, '_', *sstype*, *snum* 'EN' (for 'English') in those rows where *wnum* = 1.

The WN 2.0 to WN 3.0 mappings by Daudé et al. (2000) are plain text, in three columns (except in rare ambiguously-translated cases which have additional columns for alternatives). Table 5.5 gives an extract from the noun table:

Table 5. 5. *Sense mappings from WordNet 2.0 to 3.0 from Daudé et al. (2000)*

00001740 00001740 1
00002056 00002452 1
00002560 13740168 1
00002645 00003553 1
00003009 00004258 1
00003226 00004475 1
00004358 00005787 1
00004483 00006024 1
00004609 00006269 1
00004740 00006400 1
00004824 00006484 1
00005598 00007347 1

Joining database tables built from these plain text inputs, we produced a table named 'equivalence' (Table 5.6) which shows the mapping between Arabic WN synsets and the WordNet 3.0 identifiers used in SentiWordNet 3.0, without building a complete WordNet 3.0-aligned version of Arabic WordNet.

Table 5. 6. *Equivalence table mapping AWN synsets to WN 3.0 identifiers (irrelevant columns suppressed)*

Relation	AWN synset ID	AWN English synsetID	WN3.0 offset
equivalent	$A}ik_Aljilod_n1AR	echinoderm_n1EN	102316707
equivalent	$a>n_n1AR	concern_n1EN	105670710
equivalent	$a>n_n2AR	thing_n7EN	105855004
equivalent	$aAEa_v1AR	break_v46EN	200935987
equivalent	$aAEir_n1AR	poet_n1EN	110444194
equivalent	$aAHib_a1AR	pale_s1EN	300408992
equivalent	$aAHinap_n1AR	truck_n1EN	104490091
equivalent	$aATara_v1AR	share_v5EN	201063930
equivalent	$aATi}_AlbaHor_n1AR	seashore_n1EN	109428293

An exclusive WN 3.0 synset ID is provided by the mappings for 99.7 percent of WN 2.0 adverb synsets, 99.39 percent of noun synsets, and 98.92 percent of verb synsets from the mapped WN 2.0 synsets. Of the remaining ambiguously mapped WN2.0 synsets, a majority have no AWN linkage (Alhazmi et al., 2013).

5.2.3 Evaluation

Evaluation by human experts

We conducted a substantial evaluation, carried out by experts in the Arabic language, for scoring sentiments for each word in our ASWN database. The reason for choosing linguistic experts is because the morphology of the Arabic language is different from English; consequently, the way of expressing the language in real life is also different. Hence, the strength of sentiments in a word in Arabic and the same word in English may not be the same if we consider cross-cultural differences, which play an important role in any language. Our approach differs substantially in this regard from the method used for scoring sentiments in the SLSA (Eskander and Rambow, 2015); their scoring was done by considering scores used in the English SWN for the translations of the same word in Arabic, *which ignored the different cultural implications of the languages*. Table 5.7 shows an example of how sentiments differ for the same word in two languages.

For scoring sentiments in the ASWN, we defined three sentiment classes: positivity, negativity and neutrality. The range for scoring positivity or negativity was between 0 and 1; this follows the same scoring range used in the English SWN. In contrast, the neutrality score is determined by [1 - (positivity score + negativity score)]. Table 5.8 illustrates the statistics of ASWN and includes the number of the different part of speech (POS) tags and the percentage of the different sentiment classes.

Table 5.7. *Sentiments scores differ cross different languages*

		POS	Positivity	Negativity	Neutrality	Gloss
English	Shrivel#1	VERB	0.0	0.0	1.0	wither, as with a loss of moisture;
Arabic	Ainokama$	VERB	0.0	0.25	0.75	
English	Love#3	VERB	0.625	0.0	0.375	be enamoured or in love with; (She loves her husband)
Arabic	>aHab~a	VERB	1.0	0.0	0.0	

Table 5.8. *Statistics of ASWN with numbers of POS tags and the percentage of sentiment classes*

POS	Number	Positives	Negatives	Neutrals	Mixed
VERB	2538	22.97%	15.25%	17%	44.8%
ADJ	619	18.58%	7%	50.4%	23.9%
ADV	112	14.29%	11.6%	55.36%	18.75%
NOUN	7143	13.83%	9.6%	59.75%	16.81%

Lexicon-based classification of our corpus

The aim behind building the ASWN was to create a sentiment lexicon for this research, as until very recently, there has been no available sentiment lexicon for the Arabic language. In this section, to demonstrate the efficiency of the ASWN, we will mention briefly the results obtained for the sentiment classification phase (Phase 4 in our research framework), while a more detailed justification and discussion will take place in Chapter 9.

The ASWN was used to provide various features to train a Naïve Bayes (NB) classifier which was then applied to our corpus and obtained the following average results of F-measure: for the test dataset: for positivity 0.8048; for negativity 0.79449; hence, the average for subjectivity was 0.799; for neutrality it was 0.82098.

5.2.4 Discussion of the results

By considering the human evaluation for the sentiment task to determine the subjectivity (either positive or negative) shown in Chapter 4, Table 4.6, the average Kappa result obtained was 0.776, and the average Kappa result for facts (neutrals) was 0.789. By taking into account the results obtained after applying the ASWN, an effective improvement of the result was obtained compared to the human evaluation. In this case, the coverage of

the ASWN as a sentiment lexicon was very useful for the needs for this research. The results seen in Table 5.8 can be summarised as follows:

- The POS type NOUN represents the highest coverage in the ASWN with 7,143 words; 59.75 percent of those NOUNs were represented as neutral; 16.81 percent were mixed (i.e. could be considered either positive or negative); 13.83 percent were positive; and 9.6 percent were negative.
- The second-highest POS type coverage was for VERBs; 44.8 percent were mixed; 22.97 percent were positive; 17 percent were neutral; and 15.25 percent were negative.
- The third POS type coverage in the ASWN was for ADJs; 50.4 percent were neutral; 23.9 percent were mixed; 18.58 percent were positive; and 7 percent were negative.
- Finally, the last coverage of POS type was ADVs; 55.36 percent were neutral; 18.75 percent were mixed; 14.29 percent were positive; and 11.6 percent were negative.

By this statistic, we can conclude that the largest group of words in the ASWN were neutral, and the smallest group was negative words.

5.3 The ArTerMine tool

Extracting multi-word terms (MWTs) is an important task as far as the domain of text mining is concerned. In this section, we present an adaptation of the TerMine tool, designed to support the Arabic language: ArTerMine. The English version of the TerMine tool is a domain-independent method for the automatic extraction of MWTs (Frantzi et al., 2000). A real-data example from our corpus (Alhazmi and McNaught, 2013) is used to show how the adapted ArTerMine tool works.

5.3.1 Background: related work

TerMine, the English version, can be defined as a terminological management approach, where C-value term extraction is utilised. C-value is a domain-independent means of automatic term recognition (ATR) (Frantzi et al., 2000) using statistics and, to a lesser extent, linguistics. The linguistic element involves POS tagging and the extraction of word sequences. The statistical aspect involves attributing 'termhood' to candidate terms based on the following characteristics: (1) how often the candidate term occurs; (2) how often the candidate terms occur within other, larger terms; (3) how many of the larger terms there are; and (4) how long the candidate terms are. The statistical measure used in the C-value, as described by Frantzi et al. (2000), is shown in the following formula:

$$\boldsymbol{C\text{-}value(a)} = \begin{cases} \log_2 |a| . f(a) & if\ a\ is\ not\ nested^{22}, \\ \log_2 |a| \left(f(a) - \frac{1}{P(Ta)} \sum_{b \in Ta} f(b)\right) & otherwise \end{cases} \quad \textbf{(5. 1)}$$

Where,

- *a* represents a *candidate term*,
- *b* represents *longer candidate terms*,

[22] *Nested terms* refer to terms that occur within other longer terms (Frantzi et al., 2000).

- |*a*| represents the *length of a candidate term* (the *number of words*),
- *f(a)* represents the *frequency of occurrence* of the *candidate term a* in the corpus,
- *Ta* represents the *extracted set of candidate terms that include a*,
- *P(Ta)* represents the number of candidate terms in *Ta*,
- *f(b)* represents the *frequency of occurrence of longer candidate term b* in the corpus.

Since the TerMine tool did not support Arabic, we made an effort in this research to adapt the English version to be supportive of Arabic. This was done by applying an Arabic POS tagger to the existing TerMine tool in order to make it work on Arabic content. The Arabic POS tagger we used was built by AlGahtani (2011) who found Arabic NER complicated due to difficulties in the pre-processing stages. He found that Arabic NER is strongly linked to POS tagging. It is important to note that Arabic NER is sensitive to POS tagging errors, and that although NER is in general dependent on POS tagging performance, it is more of a problem in Arabic than in English due to the lack of capital letters as well as the high rate of potential confusion between common nouns and proper nouns (many Arabic names are synonymous with common nouns). AlGahtani (2011) concluded that, for Arabic, POS tagging and NER should not be done sequentially but rather as one step. We chose the AlGahtani (2011) tagger because of its accuracy, which was 96.6 percent on newswires, considered higher than other existing Arabic taggers (Diab et al., 2004; Mansour et al., 2007).

Regarding existing MWT extractors for Arabic, there are few that we are aware of. For instance, Attia (2008) provided an entirely linguistic technique to address Arabic MWTs. It is founded on a manually built lexicon of MWTs. The system subsequently endeavours to recognise alternative versions by employing a morphological analyser, a white space normaliser and a tokeniser. The arrangements of MWTs are detailed as trees which may be parsed to recognise the function of every element. Nonetheless, several types of MWTs are disregarded, such as, substitution compound nouns. Furthermore, the

pertinence of the extracted candidates is not calculated due to the absence of statistical measures.

Boulaknadel et al. (2008a) implemented a hybrid approach to acquire Arabic MWTs. The initial stage within their structure comprises the extraction of MWT-like units, that satisfy the following two syntactic patterns:{noun adjective, noun1 noun2} employing an available part-of-speech tagger (POS). The second stage ranks the extracted MWT-like units by employing association measures (these steps include log-likelihood ratio, FLR (Nakagawa and Mori, 2002), mutual information and t-score). The analysis procedure involves implementing the association procedures for an Arabic corpus and computing the precision of every procedure employing an amassed reference catalogue of Arabic terms.

Bounhas and Slimani (2009) used a hybrid technique to acquire multi-word terminology from Arabic corpora. For the linguistic part, they integrated two kinds of linguistic techniques detailed previously. On the one hand, they traced compound noun boundaries and recognised patterns likely to encompass compound nouns. On the other hand, they employed syntactic rules to address MWTs. These rules are founded on linguistic data: a morphological analyser and a POS tagger. For the numerical part they implemented the Log-Likelihood Ratio (LLR) method. During the analysis stage, they employed a similar corpus and reference catalogue to that which had been employed in Boulaknadel et al. (2008a). Their findings held promise, particularly with MWTs (Bounhas and Slimani, 2009).

An alternative approach has been suggested by El-Khatib et al. (2010) for the extraction of MWTs from an Arabic corpus. Their focus was compound nouns as a significant kind of MWT and on suitable bigram terms. The technique is dependent on two filters: (1) a linguistic filter, in which new syntactic patterns are suggested based on definite and indefinite kinds of nouns; the extraction of the candidate MWTs is based on

the pattern of nouns, in addition to patterns of nouns which are linked by a preposition; and (2) a statistical filter, within which the Unithood measure was taken into account through selecting an Log-Likelihood Ratio (LLR) measure as it provides appropriate outcomes for the extraction of Arabic MWTs (Bounhas and Slimani, 2009). Regarding termhood, they implemented a C-value measure as it has been broadly acknowledged as a valuable technique for the ranking of candidate MWTs. The LLR method may be employed effectively as an indication of association measure for the two words within a bigram (El-Khatib and Badarenh, 2010).

5.3.2 Applying ArTerMine on the research corpus

The The aim behind the adaptation of the ArTerMine from the English TerMine tool was threefold: (1) the availability of TerMine[23]; we have a licence to access and adapt the code; (2) the availability and accuracy of AlGahtani's (2011) tagger, which made its use worthwhile; and (3) our aim to have an Arabic version of TerMine for future work. In order to check the performance of the ArTerMine tool, random samples of different blog posts from our corpus were used. The maximum length strings we extracted consisted of 10 words, which were aimed at having long MWTs. Table 5.9, illustrates a sample of our corpus, shown with tagged and tokenised data, which is the input to the ArTerMine tool. Moreover, a sample of the ArTerMine output containing MWTs with transliterations is shown in Table 5.10.

[23] http://www.nactem.ac.uk/software/termine/

Table 5.9. *A sample of the corpus*

Original text in Arabic
Arcaptcha : تقنية الكابتشا باللغة العربية مع بعض الإصرار والتعب والجد والجهد قام الشابين عبدالعزيز المقري و ناصر الوهيبي من جامعة البترول و المعادن بإطلاق مشروع عربي هو الأول من نوعه في عالمنا . وهي عبارة تطوير لتقنية كاباتشا من خلال دعمها للغة العربيه وتمت تسميتها بـ Arcaptcha. وهو مكتبة برمجية لمطوري المواقع و الانظمة المتصلة بالشبكة للتصدي للطلبات الوهمية والاوتوماتيكية . حيث أن الكابتشا هو اختبار يميز بين إجابة المستخدم الأدمي وبرامج الحاسوب. ولكن العديد منا أصابه الملل من الحروف اللاتينية التي كنا نحن كعرب مجبرين على استخدامها كل مرة. فاللغة هي وعاء حضارتنا. ومن هذا المنطلق أتت فكرة المشروع الجديد بهدف تكثيف المحتوى العربي و تسهيل إيصاله لجميع المستخدمين العرب. وأنا شخصيا أدعم هذه الفكرة و أؤيدها بشده . كما أنه هذا وهو متوفر الأن للتحميل والإطلاع أو تجربة البرنامج من خلال المتصفح .
After the tokenisation and tagging
Arcaptcha/NNP :/PUNC tqnyp/JJ AlkAbt$A/NNP b/IN+Allgp/NN AlErbyp/JJ mE/IN bED/NN Al w/CC+AltEb/NN w/CC+Aljd/NN w/CC+Aljhd/NN qAm/VBD Al$Abyn/NNS EbdAlEzyz/NNP AlmqrY/NN w/CC nASr/NNP Alwhyby/NNP mn/IN jAmEp/NN Albtrwl/NN w/CC AlmEAdn/NN b/IN+ m$rwE/NN Erby/JJ hw/PRP Al>wl/JJ mn/IN nwE/VBD+h/PRP fy/IN EAlm/JJ+nA/PRP$ ./PUNC why/VBD EbArp/NN tTwyr/NN l/IN+tqnyp/NN kAbAt$A/NNP mn/IN xlAl/NN dEmhA/NN l/IN+Allgp/NN AlErbyp/JJ w/CC+tmt/VBP tsmyp/NN+hA/PRP$ b/IN Arcaptcha./NNP w/CC+hw/NNP mktbp/NN brmjyp/JJ l/IN+mTwry/JJ+y/PRP$ AlmwAqE/NN w/CC AlAnZmp/NN AlmtSlp/JJ b/IN+Al$bkp/NN l/IN+AltSdy/NN l/IN+AlTlbAt/NNS Alwhmyp/JJ w/CC+Al>wtwmAtykyp/JJ ./PUNC Hyv/WRB >n/IN AlkAbt$A/NNP hw/PRP AxtbAr/NN ymyz/VBP byn/IN Almstxdm/JJ Al|dmy/JJ w/CC+brAmj/NN AlHAswb/NN ./PUNC w/CC+lkn/CC AlEdyd/NN mnA/NNP >SAb/VBD+h/PRP Alml1/NN mn/IN AlHrwf/NN AllAtynyp/JJ Alty/WP knnA/VBD nHn/PRP k/IN+Erb/NN mjbryn/NNS ElY/IN AstxdAmhA/NN kl/NN mrp/NN ./PUNC f/CC+Allgp/NN hy/PRP wEA'/NN HDArp/NN+nA/PRP$ ./PUNC w/CC+mn/IN h*A/DT AlmnTlq/NN >tt/VBD fkrp/NN Alm$rwE/NN Aljdyd/JJ b/IN+hdf/NN tkvyf/NN AlmHtwY/NN AlErby/JJ w/CC tshyl/NN l/RP+jmyE/NN Almstxdmyn/NNS AlErb/JJ ./PUNC w/CC+>nA/PRP $xSyA/JJ >dEm/VBP h*h/DT Alfkrp/NN w/CC >&yd/VBP+hA/PRP b/IN+$d/NN+h/PRP$ kmA/IN >n/IN+h/PRP h*A/DT w/CC+hw/NNP mtwfr/JJ Al>n/RB l/IN+AltHmyl/NN w/CC+Al{TlAE/NN >w/CC tjrbp/NN AlbrnAmj/NN mn/IN xlAl/NN AlmtSfH/NNP ./PUNC

Table 5.10. *A sample of ArTerMine output*

Multi-words terms
mktbp brmjyp mTwry AlmwAqE Al>nZmp AlmtSlp Al$bkp AltSdy AlTlbAt Alwhmyp Al>wtwmAtykyp
fkrp Alm$rwE Aljdyd hdf tkvyf AlmHtwY AlErby
jAmEp Albtrwl AlmEAdn <TlAq m$rwE Erby
Al$Abyn EbdAlEzyz AlmqrY nASr Alwhyby
<jAbp Almstxdm Al|dmy brAmj AlHAswb
EbArp tTwyr tqnyp kAbAt$A
Arcaptcha
dEmhA Allgp AlErbyp
Almstxdmyn AlErb
tshyl <ySAl
AlHrwf AllAtynyp
Ymyz
>dEm
>&yd
AlEdyd mnA >SAbh Alml1 mn AlHrwf AllAtynyp Alty knA nHn kErb mjbryn ElY AstxdAmhA kl mrp
tshyl <ySAlh ljmyE Almstxdmyn AlErb
>dEm h*h Alfkrp w >&ydhA b$dh

5.3.3 Evaluation

The ArTerMine tool is generally based on the C-value measure whose scores for candidate MWTs represent a complex calculation of significance rather than straightforward frequency of occurrence (Dagan and Church, 1995; Justeson and Katz, 1995). However, we evaluated the results of ArTerMine in terms of precision and recall based on experts' annotation agreement. That is, we focused on whether the extracted MWTs matched the terms annotated by experts, and *not on the specific C-value*[24] *scores*. Figure 5.2 and Figure 5.3 illustrate examples from our corpus (a) expert annotations and (b) ArTerMine annotations; the full results[25] extracted by the ArTerMine tool are shown in Appendix A.

Arcaptcha : تقنية الكابتشا باللغة العربية
مع بعض الإصرار والتعب والجد والجهد قام الشابين عبدالعزيز المقري و ناصر الوهيبي من جامعة البترول و المعادن بإطلاق مشروع عربي هو الأول من نوعه في عالمنا . وهي عبارة تطوير لتقنية كاباتشا من خلال دعمها للغة العربيه وتمت تسميتها بـ Arcaptcha. وهو مكتبة برمجية لمطوري المواقع و الانظمة المتصلة بالشبكة للتصدي للطلبات الوهمية والاوتوماتيكية . حيث أن الكابتشا هو اختبار يميز بين إجابة المستخدم الأدمي وبرامج الحاسوب. ولكن العديد منا أصابه الملل من الحروف اللاتينية التي كنا نحن كعرب مجبرين على استخدامها كل مرة. فاللغة هي وعاء حضارتنا. ومن هذا المنطلق أتت فكرة المشروع الجديد بهدف تكثيف المحتوى العربي و تسهيل إيصاله لجميع المستخدمين العرب. وأنا شخصيا أدعم هذه الفكرة و أؤيدها بشده كما أنه هذا وهو متوفر الأن للتحميل والإطلاع أو تجربة البرنامج من خلال المتصفح .

Figure 5. 2. *A sample of experts' annotations (in yellow)*

Arcaptcha : تقنية الكابتشا باللغة العربية
مع بعض الإصرار والتعب والجد والجهد قام الشابين عبدالعزيز المقري و ناصر الوهيبي من جامعة البترول و المعادن بإطلاق مشروع عربي هو الأول من نوعه في عالمنا . وهي عبارة تطوير لتقنية كاباتشا من خلال دعمها للغة العربيه وتمت تسميتها بـ Arcaptcha. وهو مكتبة برمجية لمطوري المواقع و الانظمة المتصلة بالشبكة للتصدي للطلبات الوهمية والاوتوماتيكية . حيث أن الكابتشا هو اختبار يميز بين إجابة المستخدم الأدمي وبرامج الحاسوب. ولكن العديد منا أصابه الملل من الحروف اللاتينية التي كنا نحن كعرب مجبرين على استخدامها كل مرة. فاللغة هي وعاء حضارتنا. ومن هذا المنطلق أتت فكرة المشروع الجديد بهدف تكثيف المحتوى العربي و تسهيل إيصاله لجميع المستخدمين العرب. وأنا شخصيا أدعم هذه الفكرة و أؤيدها بشده كما أنه هذا وهو متوفر الأن للتحميل والإطلاع أو تجربة البرنامج من خلال المتصفح .

Figure 5. 3. *A sample of ArTerMine annotations (in green)*

[24] For more information about the original TerMine and the nature of C-value, see (Frantzi et al. 2000).
[25] Refer to Table 5.4 for a sample of the results, or to Appendix A for the entire list of terms extracted with C-value.

The results obtained from ArTerMine yielded an F-measure score of 0.81 with 0.84 recall and 0.78 precision.

5.3.4 Discussion of the results

We noticed that most of the terms annotated by experts were extracted by the ArTerMine tool. For our purposes, as we are looking for long terms and ignoring frequency, we consider the F-measure of 0.81 to be a good result for MWT extraction (Attia, 2008; Bounhas and Slimani, 2009; El-Khatib and Badarenh, 2010; Meryem et al., 2014). Although this work was tested only on our corpus, since the term extraction technique is domain independent, we are confident the ArTerMine tool will be useful for other Arabic MWT extraction applications.

5.4 TechTerms list

Having a list that includes technological terms and company names, written either in English or in Arabic, or transliterated from English to Arabic, was one of our research contributions (RC5). The aim behind having this list was to provide our corpus with a lexicon that could support the needs for extracting those TechTerms from our corpus. Considering those TechTerms was one of our corpus challenges, as shown in Chapter 4, Table 4.2.

5.4.1 Preparation and requirements

In order to build the TechTerms list, we took into account the following steps: (1) the manual work required to build the resource, (2) designing all the requirements needed to be included in the resource.

The manual work required

During the preparation step, a substantial amount of manual work was undertaken, this involved gathering information representing the names of any kind of (technology, software, devices, …, etc.), and/or names of companies for selling/producing these kinds of technologies. As our focus when compiling the TechTerms list was on two languages (English and Arabic), we attempted to include all technological terms and/or company names for technology in both languages.

Moreover, these technological terms were collected from (1) our research corpus, by considering all the terms that represent any type of technologies/companies names either in (English or Arabic) to be included; we detected that 26% of the technological terms in the TechTerms list were collected from the research corpus; (2) we searched on the Internet for names of famous companies, and forms of technology such as software, applications and devices, etc., in order to achieve a wider coverage by the TechTerms list;

we detected that 74% of technological terms were collected from *several lists*[26] from the Internet.

Designing the requirements

The next step after gathering all technological terms required to build the TechTerms list, was to design all the requirements that needed to be included in the list, as shown in Figure 5.4. This led us to identify four main classes of words. In the figure, English Companies and English Technologies refer to company names or technologies written in the Roman alphabet; and Arabic Companies and Arabic Technologies refer to such names/technologies written in Arabic.

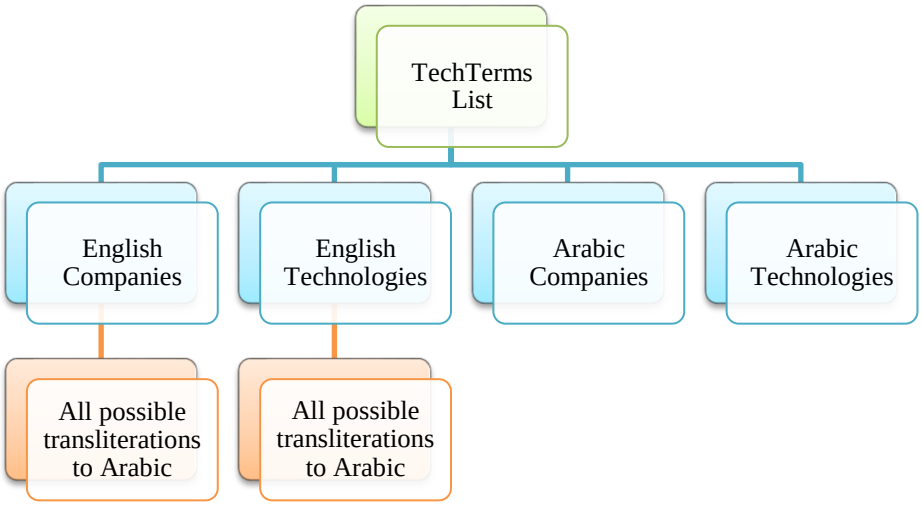


Figure 5. 4. *Classes required for designing the TechTerms list*

To date, the total number of entries on the TechTerms is about 8,700. As we are trying to make it an open resource for forms of technology in English and Arabic, we will keep updating[27] the list with any new companies or technology that arise. Table 5.11 shows examples of each class in the TechTerms list.

[26] Details of lists collected from the Internet are provided in Appendix B.
[27] Because this was not a core part of the research, in future we would rather use/train a NER tool to help with such updating.

Table 5.11. *Examples of each class in the TechTerms list*

English Companies	Transliterations to Arabic	English Technologies	Transliterations to Arabic	Arabic Companies	Arabic Technologies
Microsoft	مايكروسوفت	iPhone	ايفون	حسوب	مغردون
	ميكروسوفت		آي فون	المحترفون العرب	الكابتشا العربية
Apple	أبل	Gmail	جي ميل	بديل	كويتي ريدر
	آبل		جيميل	المعلومات العالمية	سعودي ريدر
	ابل		ج ميل	الحلول المتكاملة	سمة
Samsung	سامسونج	Nexus	نيكسوس	الدار العربية للتقنية	اقرأ مع مومو
	سامسونغ		نيكسوز	تاكو للألعاب	حاسبة الزكاة

By comparing ours with the existing lists on the Internet, we achieved better and a wider coverage for the TechTerms list for three important reasons:

1 The existing lists concentrated on one language only, either English or Arabic, while our list considered both languages.

2 Furthermore, these existing lists represented either forms of technologies or company names whereas our list included both.

3 The TechTerms list was differentiated from the existing lists by its coverage and inclusion of the transliterated forms for company names or technologies written in the Roman alphabet.

5.4.2 Evaluation and discussion of the results

This section will mention briefly the results, whereas in Chapter 9 will go into more detail. In order to evaluate the effectiveness of the TechTerms list, we applied it to our corpus, as a lexicon-based method, to determine all the technological terms. We recorded an F-score of 0.851 for the test datasets. As can be seen, the outcomes of the results met our needs when building this TechTerms list.

5.5 Summary

We expounded in the previous chapter – Chapter 4 – the motivation for creating the TechTerms list and the ASWN lexicon, and for the adaptation of text mining tools to extract the multi-word terms. Therefore, in this chapter we talked about the issues pertaining to the creation of such lexicons, as well as the ArTermine tool, an extension to the English language version of TerMine.

In this study, the corpus data was retrieved from technology blogs and relevant (technology-based) tweets from Twitter, and resources just described were designed to be representative of the domain of the corpus. The next step is assess whether the extracted data is objective and subjective, and whether the latter has any polarity, i.e. positive, negative or simply neutral.

The SentiWordNet lexicon (SWN) is English language-based, and complements the English WordNet (WN), with statistical indications of the sentiment expressed by each word sense. Our literature review did not find a relevant Arabic version suitable for our purposes, and therefore we proceeded to create Arabic SWN (ASWN). This meant upgrading the existing Arabic WN version 2, to the Arabic WN version 3; then, mapping the English SWN3 to the Arabic WN3; and finally appraising and amending the synsets of the Arabic WN3, using the English SWN3 to be maintained in the ASWN. This resulted in a 10,500 word database that comprises nouns, verbs, adjectives and adverbs.

As noted earlier, there were other Arabic-based tools that nearly serve the same aim as our research, but not close enough to utilise them. In particular, the sentiment lexicon for standard Arabic, SLSA, could have been substituted for ASWN, had it become available earlier, however, its lack of attention to cross-cultural sentiment aspects would nevertheless argue against its use. In contrast, the sentiment annotations in ASWN (each polarity in a scale of 0 to 1) were both created and evaluated by Arabic language experts, and thus stand to reflect more faithfully any cultural differences from English.

Using the F-measure, the results for the test dataset were: for positivity 0.805; for negativity 0.794; hence, the average for subjectivity was 0.799; and finally, for neutrality it was 0.821.

As for Multi-Word Terms (MWTs), as TerMine (an already available and accurate tool based on the C-value technique) did not support Arabic, therefore we adapted the English version to support Arabic (ArTerMine), by incorporating an Arabic POS tagger. We evaluated the comparison between extracted MWTs of the experts and the extracted MWTs by ArTerMine, and found the F-score was 0.81, made up of 0.84 recall and 0.78 precision.

In order to create our open-list of technology terms (TechTerms list), we set out guidelines for inclusion in the list. This included the four main classes of words: English companies, English technologies, Arabic companies and Arabic technologies, supplemented by alternative transliterations of the English terms. The creating process was based on our own knowledge of the subject terms, as well as carrying out search procedures from the Internet. This resulted in an 8703-entry list. We tested the efficiency of this list, by applying it to our corpus, and found that the F-score was 0.851 for the test dataset.

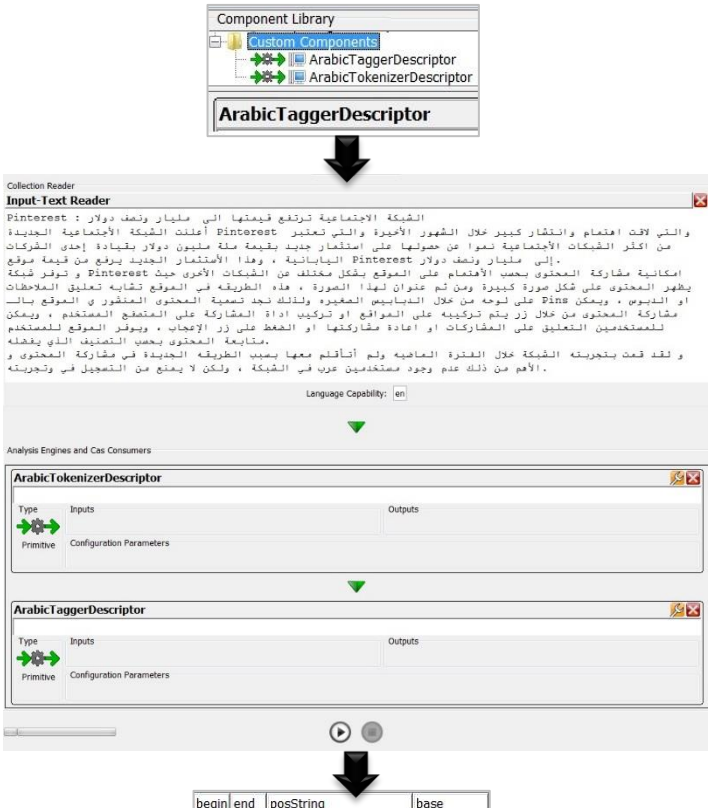
CHAPTER 6: CORPUS DATA PROCESSING

In this chapter, we describe in detail Phase 1 of our research framework. We outline the processes that our corpus went through in order to be analysed and prepared for the remaining phases included in the research framework, as shown in Figure 1.2. We started with the raw data from the corpus and ended up with MWTs extracted by ArTerMine. We made use of U-Compare[28]in this phase.

U-Compare is a system that integrates text mining and NLP, relying on the Apache UIMA framework; it does this by providing access to a large set of ready-to-use interoperable NLP components (Kano et al., 2011; Kano et al., 2009). U-Compare is currently accessible as a repository for the largest number of UIMA-based text mining components (Ananiadou et al., 2011; Thompson et al., 2011). Furthermore, U-Compare has a simple and an easy interface that allows users to generate complex NLP workflows by dragging and dropping each component to make the outputs of these workflows reachable and simple through comparison and visualisation (Kolluru et al., 2011; Kontonatsios et al., 2011).

[28] http://nactem.ac.uk/ucompare/

6.1 Data processing



begin	end	posString	base
0	4	qymp/NN	qymp
5	9	$bkp/NN	$bkp
10	19	Pnterest/NNP	Pnterest
20	30	Al>jtmAEyp/JJ	Al>jtmAEyp
31	36	trtfE/VBP	trtfE
37	40	AlY/IN	AlY
41	44	1.5/CD	1.5
45	50	mlyAr/NN	mlyAr
51	56	dwlAr/NN	dwlAr
56	57	./PUNC	.
58	63	>Elnt/VBD	>Elnt
64	70	Al$bkp/NN	Al$bkp
71	81	Al>jtmAEyp/JJ	Al>jtmAEyp
82	89	Aljdydp/JJ	Aljdydp
90	99	Pnterest/NNP	Pnterest
100	105	w/CC+Alty/WP	wAlty
106	110	lAqt/VBD	lAqt
111	117	AhtmAm/NN	AhtmAm
118	125	w/CC+{nt$Ar/NN	wAnt$Ar
126	130	kbyr/JJ	kbyr
131	135	xlAl/IN	xlAl
136	142	Al$hwr/NN	Al$hwr
143	150	Al>xyrp/JJ	Al>xyrp
151	156	w/CC+Alty/WP	wAlty
157	162	tEtbr/VBN	tEtbr
163	165	mn/IN	mn
166	170	Akvr/JJ	Akvr
171	178	Al$bkAt/NNS	Al$bkAt
179	189	Al>jtmAEyp/JJ	Al>jtmAEyp
190	194	nmwA/NN	nmwA
195	197	En/IN	En
198	204	HSwlhA/NN	HSwlhA
205	208	ElY/IN	ElY
209	216	AstvmAr/NN	AstvmAr
217	221	jdyd/JJ	jdyd
222	227	b/IN+qymp/NN	bqymp
228	231	m}p/CD	m}p
232	237	mlywn/CD	mlywn
238	243	dwlAr/NN	dwlAr
244	250	b/IN+qyAdp/NN	bqyAdp
251	255	<HdY/NN	<HdY
256	263	Al$rkAt/NNS	Al$rkAt
264	273	AlyAbAnyp/JJ	AlyAbAnyp

Figure 6.1. *The U-Compare workflow using the Arabic POS tagger*

After generating our corpus from Arabic technology blogs, we removed the XML tags, and then we dealt with each blog post separately in analysis according to the remaining phases in the research framework. Next, in order to provide a valuable corpus for the domain of Arabic opinion mining and sentiment analysis, spelling mistakes were checked and corrected manually (because dealing with spelling errors was outside the scope of the research) as shown in Phase 1: corpus data processing (Figure 1.2). Then, our corpus was ready to be run through AlGahtani's (2011) POS tagger, which we configured as a component in U-Compare. Figure 6.1 illustrates the U-Compare workflow until we reach our tagged data.

An Arabic blog post is shown as input in the illustration. Then, the POS tagger was applied. Finally, we ended up with the tagged data—the workflow output. All files were saved as (.XMI[29]), which is the acceptable format for ArTerMine.

The next step is to run the data through ArTerMine, which was available for research purposes under the NaCTeM[30] licence. Figure 6.2 shows the ArTerMine interface.

Figure 6.2. *ArTerMine interface*

On providing the NaCTeM access key, the user then selects the mode of MWTs to be extracted. For more information about Quick and Full modes, see Frantzi et al. (2000)

[29] XML Metadata Interchange (XMI) Specification, Version 2.0. http://www.omg.org/docs/formal/03-05-02.pdf.
[30] http://www.nactem.ac.uk/

and Okazaki and Ananiadou (2006). For the purpose of our research, we used the Quick mode as it gave us useful MWTs which met our needs.

6.2 Summary

This chapter introduced the first phase of our proposed research framework, which is the initial process that our corpus went through. We started with the raw Arabic texts and prepared them by removing XML tags and correcting spelling mistakes manually. Then, we used the Arabic POS tagger via the U-Compare workflow. Finally, ArTerMine was applied to extract MWTs for each blog post.

CHAPTER 7: CLUSTERING OF BLOG POSTS AND SIMILARITY

In this chapter, we focus on Phase 2 of our research framework, as shown in Figure 1.2. A hybrid clustering method was used in this step; this is considered one of our research contributions (RC_6). In this thesis, a "cluster" means a group of words and/or multi-word phrases, extracted from a text such as a blog post; and "clustering" means extracting significant words and phrases from a text.

Phase 2 was carried out using LDA topic modelling offered by MALLET by combining and integrating three levels of clustering: (1) by using only the raw text in a blog post as the first level to yield clusters of single words; (2) then, using only MWTs related to this blog post as the second level to yield clusters of MWTs; and (3) then applying both together as the third level to yield clusters including both single words and MWTs. From the outcome of this hybrid method, we wanted to demonstrate the sufficiency of integrating Arabic MWTs for the purpose of obtaining both meaningful and useful clusters that reflected the same theme as the blog post.

The next step in this phase was the application of a similarity filter to evaluate the final collection of clusters from our hybrid method, the three levels; we made use of the Jaccard coefficient and the Cosine similarity measure in this step. The reason for using the similarity filter here was to obtain a collection of useful clusters that could be used for further processes in the framework (in Phase 4; see Chapter 9). Finally, the collection of similar clusters was evaluated against a gold standard of clusters made by human experts.

7.1 Background: related work

A number of studies have been carried out to categorise associated information and to sustain the management of available texts on the Internet. The most widespread method employed in classifying documents, which include associated information in one collection or a group, is called document clustering (Froud et al., 2013). This method enhances the document distribution procedure with data about similarity among documents. For the proficient arrangement of documents, document clustering comprises a basic and enabling instrument (Mccallum, 2002; Řehůřek et al., 2011; Rosen-zvi et al., 2004).

Clustering is the procedure of gathering items portrayed in a similar state within even divisions (clusters). Within document clustering, the items considered are texts. The requirement for this grouping is clarified by the great number of texts that are usually found within a collection of documents. Lu et al. (2011) state that two approaches to document clustering are normally taken, both of which employ topic models. The first approach employs a topic model to decrease the dimension of portrayal of documents (from word representation to topic representation) and subsequently, to implement a standard representation algorithm for the new representation; the second approach employs topic models more immediately.

Regarding the character of the Arabic language, the writing structure, the direction of the writing, the absence of vowels and the morphological arrangement of the language have all limited the number of studies on this language, particularly regarding automatic classification (clustering or categorisation). The majority of studies in the literature concentrate on the morphological dimension of the language (Ababneh et al., 2012; Larkey et al., 2007) through establishing pre-processing instruments like stemming and showing their effect on IR or on supervised classification (categorisation). A small number of studies concentrate on document clustering (Alghamdi et al., 2014; Kelaiaia

and Merouani, 2016). For instance, there has been observation of the present methods of Web page evaluation, which vary in line with the employed levels of classification (phrase level, sentence level or document level) or the kinds of features contemplated for the employed methods. Abbasi et al. (2008a) state that the kinds of features noted are (a) syntactic, which is related to the word structure where meaning of the word is taken into consideration; and (b) stylistic, in which the focus is on the style of the word or phrase (Abbasi et al., 2008a).

One study on sentiment analysis that has been conducted with documents in the Arabic language is by Farra et al. (2010). The morphology of Arabic presents challenges, according to Saleh and Al-Khalifa (2009) and Beseiso et al. (2011) who contend that establishing a structure that allows the Arabic language to be understood and processed by machines requires specific and specialised processes. Farra et al. (2010) propose that there are two levels to Arabic text sentiment mining, the document level and the sentence level. In their research, they employed the identified polarities of the sentences to categorise the overall polarity of the document (Farra et al., 2010).

Syntactic and stylistic features were employed concurrently by Abbasi et al. (2008a) to classify the opinions within multilingual (English and Arabic) Web forums. Semantic features are not taken into consideration within the categorisation procedure. The effects of stemming on the Arabic text document clustering were explored by Froud et al. (2010). Their research established that the representation of the documents, as well as pre-processing, may reduce the document features and hasten the clustering.

Other studies have concentrated on approaches that can be employed to classify documents in line with semantic similarities, although with languages other than Arabic. Shaban (2009) suggests a technique for clustering documents in line with semantic data through establishing the similarities within documents. This technique uses the semantic components to offer a measure of similarity that is precise. The technique may be

employed to address challenges of document clustering. Ultimately, efficient document clustering is provided by the technique which is capable of identifying the structures and meaning of text within documents (Shaban, 2009).

In our research, the most suitable approach was to employ clustering not amid various issues in various documents, but for one issue within a blog post, with its associated MWTs from ArTeMine. MALLET LDA topic modelling was used, as it offers a simplified way of analysing great amounts of unlabelled text. A 'topic' is a cluster of words which regularly occur simultaneously. Topic modelling may connect words with comparable meanings and distinguish between utilises of words that have various meanings employing contextual clues (Steyvers and Griffiths, 2007).

7.2 Method

We hypothesised that applying MWTs in the clustering process with their related blog posts would easily help us to focus on and categorise information based on important content. By using such clustering, we hoped to identify information that would reflect the same issue that the original blog post was talking about, in order to provide us with a collection of meaningful and useful material.

Our method consisted of two stages. We used MALLET LDA topic modelling clustering approach, as illustrated in Figure 7.1:

1 In the first stage, clustering was performed with each level; each 'class' was treated separately, using the whole blog post (class 1), related MWTs (class 2) and then the blog post together with its MWTs (class 3). Then, the outcome of the three classes was combined to build a collection of clusters.

2 In the second stage, a similarity filter was used to determine a collection of useful clusters.

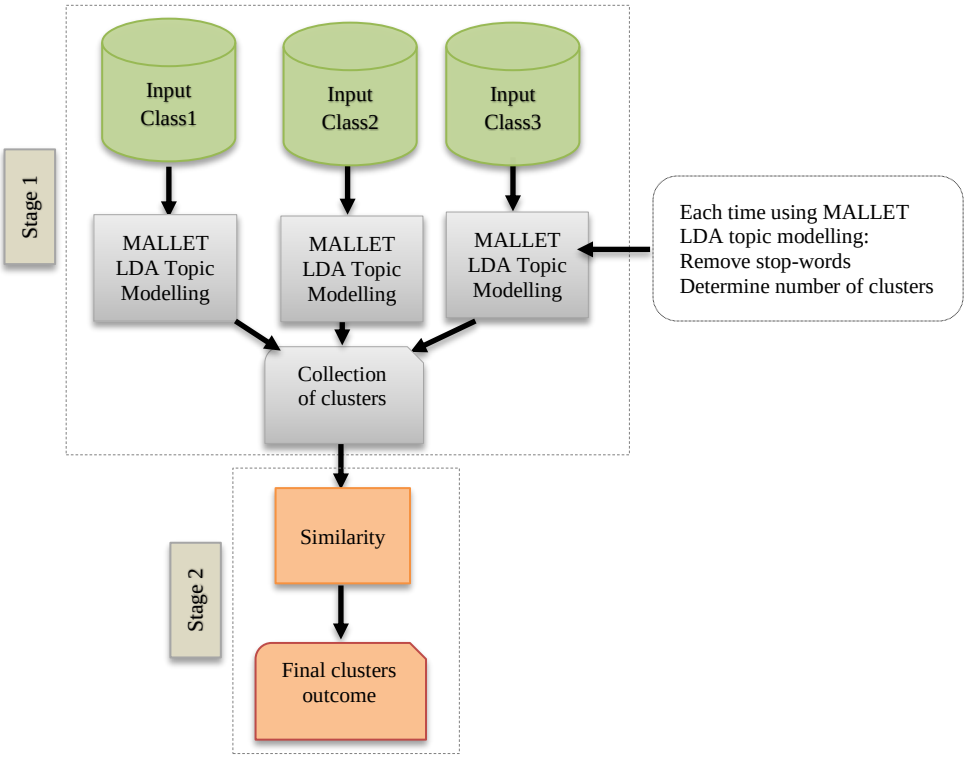


Figure 7.1. *Two-stage method for clustering*

7.2.1 Stage 1: MALLET LDA topic modelling for clustering

MALLET provides a particular input command to convert the data into MALLET's specific internal format (McCallum, 2002; McCallum et al., 2005). The LDA (Latent Dirichlet Allocation), since its introduction by Blei et al. (2003), has attracted considerable attention and interest from the statistical ML (machine learning) and NLP (natural language processing) communities (Kelaiaia and Merouani, 2016). LDA was defined by Blei et al. (2003) as "a generative probabilistic model of a corpus" and the basic idea behind the LDA is straightforward topic modelling, which means that *documents contain a mixture/multiple topics* and each topic can be characterised as a *distribution over fixed words* (Blei et al., 2003; Alghamdi and Selamat, 2015).

Befor starting with our method at this stage, in order to understand the mathematical notation behind the LDA, as described by Steyvers and Griffiths (2007), assume we have:

1. $P(z)$ signifies the distribution across topics z within a specific document.
2. $P(w \mid z)$ signifies the probability distribution across words w given topic z, let us call it a topic-word distribution.
3. Every word w_i in a document is the product of sampling one of the topics in the topic distribution, and subsequently selecting a word from the topic-word distribution.
4. $P(z_i = j)$ represents the probability of the j topic having been sampled for the i word and $P(w_i \mid z_i = j)$ being the probability of word w_i in topic j.

   Therefore, the distribution over words in a document is set out by the model:

$$\boldsymbol{P}(\boldsymbol{w_i}) = \sum_{\boldsymbol{j=1}}^{\boldsymbol{T}} \boldsymbol{P}(\boldsymbol{w_i} \mid \boldsymbol{z_i} = \boldsymbol{j})\, \boldsymbol{P}(\boldsymbol{z_i} = \boldsymbol{j}) \qquad \textbf{(7. 1)}$$

In which T represents the number of topics. For the sake of simplicity, assume that $\varphi (j) = P(w \mid z=j)$ represents the multinomial distribution across words for topic j and $\theta (d) = P(z)$ denotes the multinomial distribution across topics for document d.

In addition, suppose that the text group/collection includes D documents, each d of these documents consisting of Nd words. Assume N represents the total numbers of words (i.e., $N = \Sigma Nd$). The parameters φ and θ designate the important words for each topic, and also the important topics for each document, correspondingly.

Blei et al. (2003) defined the probability density of a T dimensional Dirichlet (Dir) distribution across the multinomial distribution $p=(p1, \ldots, pT)$ as follows:

$$\boldsymbol{Dir}(\boldsymbol{\alpha_1}, \ldots, \boldsymbol{\alpha_T}) = \frac{\Gamma(\sum_j \boldsymbol{\alpha_j})}{\prod_j \Gamma(\boldsymbol{\alpha_j})} \prod_{j=1}^{T} \boldsymbol{P_j}^{\boldsymbol{\alpha_j - 1}} \qquad \textbf{(7. 2)}$$

The parameters of the above distribution are indicated by $\alpha_1 \ldots \alpha_T$. Each parameter αj are viewable as a prior observation calculation of the total times topic j is sampled in a document, prior to actually observing any words within the document.

Now, we conducted a method of topic modelling using the MALLET LDA topic modelling. At this stage, we removed the stop-words from the input data. Once the data was formatted and available to the LDA implementation offered by MALLET, it was time to trial the topic modelling on our input data.

Additionally, we determined the number of topics we wanted to be clustered; by 'determining the number of topics' here we mean specifying and identifying the amount of different kinds of information within a blog post, for example, types of technologies mentioned, definitions of these technologies, opinions about them, and so on.

After running our experiments on the three classes several times, we determined that the number of clusters should be 30 for long blog posts and 14 for short ones; this provided us with the number of clusters that we were looking for, as the blog posts varied in length from eight to 298 sentences. Figure 7.2 shows a sample of MALLET LDA topic modelling output.

```
 0      0.5     >tt fkrp Alm$rwE Aljdyd bhdf tkvyf AlmHtwY AlErby w tshyl <ySAlh
ljmyE Almstxdmyn AlErb
 1      0.5     AlEdyd mnA >SAbh Alml1 mn AlHrwf AllAtynyp Alty knA nHn kErb
mjbryn ElY AstxdAmhA kl mrp
 2      0.5     mktbp brmjyp lmTwry AlmwAqE w AlAnZmp AlmtSlp bAl$bkp lltSdy
llTlbAt Alwhmyp w AlAwtwmAtykyp
 3      0.5     tTwyr ltqnyp kAbAt$A mn xlAl dEmhA llgp AlErbyh wtmt tsmythA b
Arcptch
 4      0.5     AxtbAr ymyz byn <jAbp Almstxdm Al>dmy w brAmj AlHAswb
 5      0.5     tqnyp AlkAbt$A
 6      0.5     Arcptch  tqnyp AlkAbt$A b Allgp AlErbyp
 7      0.5     m$rwE Erby
 8      0.5     Al>wl mn nwEh fy EAlmnA
 9      0.5     >dEm h*h Alfkrp w >&ydhA b$dh
```

Figure 7.2. *A sample of MALLET LDA topic modelling output; the columns represent: (i) the topic nmber (ii) Dirichlet parameter for the topic (which is a default value so this is why every topic in this output has the number 0.5) (iii) clusters.*

As a result of our hybrid method, we noticed that the use of MWTs (class 2) provided us with short, useful clusters (as defined by the similarity measures used, see section 7.2.2), for example, the name of the technology that the blog post referred to. Figure 7.2 contains an example of a cluster:

```
tqnyp AlkAbt$A
```

which represents the name of an Arabic form of technology. It is important to mention that our hybrid clustering algorithm gives us soft (overlapping) clusters, i.e. the same cluster mentioned above; this string also occurs in cluster 6 as shown in Figure 7.2. Furthermore, through the use of class 1, the actual text within a blog post, and then with its MWTs in class 3, we were provided with some long clusters that included, for example, a definition of this technology:

```
mktbp brmjyp lmTwry AlmwAqE w AlAnZmp AlmtSlp bAl$bkp
lltSdy llTlbAt Alwhmyp w AlAwtwmAtykyp
```

or an opinion about the technology:

```
>dEm h*h Alfkrp w >&ydhA b$dh
```

Thus, by applying the three level clustering of each blog post in our corpus as shown above, we identified the following:

1 What technology that the blog post talking about.

2 Related information regarding this technology.

3 Opinions about this technology.

So, from this, we can say that our generated clusters for each blog post represented and determined information about the content of that particular blog post.

After building the collection of clusters from Stage 1, the next step was to go through the second stage of this process, the similarity filter, in order to validate and identity a collection of useful clusters only for each blog post in our corpus.

7.2.2 Stage 2: similarity filter

To measure the quality of the whole collection of clusters for each blog post, we used the similarity filter here between clusters and the original sentences within the blog post. This was done by using the Jaccard coefficient and the Cosine for measuring similarity.

The Jaccard coefficient

The Jaccard coefficient is considered a classic statistical measurement for similarity in a group or a collection (Jaccard, 1901; Niwattanakul et al., 2013; Schluter and Harris, 2006). It is defined as shown in formula 7.3:

$$\boldsymbol{Jaccard}\ (\boldsymbol{A}, \boldsymbol{B}) = \frac{(\boldsymbol{A} \cap \boldsymbol{B})}{(\boldsymbol{A} \cup \boldsymbol{B})} \qquad \textbf{(7. 3)}$$

The Jaccard coefficient's range is between 0 and 1, and it can be computed in our case by considering $\boldsymbol{A}$ to be clusters and $\boldsymbol{B}$ to be sentences within a blog post. By this, we mean that the intersection between a cluster and a sentence is computed based on common words; furthermore, the union of all words in both of them is computed. Thus, we computed the Jaccard coefficient using the following formula 7.4:

$$\boldsymbol{Jaccard}\ (\boldsymbol{cluster}, \boldsymbol{sentence}) = \frac{(\boldsymbol{words\ in\ a\ cluster}) \cap (\boldsymbol{words\ in\ a\ sentence})}{(\boldsymbol{words\ in\ a\ cluster}) \cup (\boldsymbol{words\ in\ a\ sentence})} \qquad \textbf{(7. 4)}$$

Hence, for each collection of clusters, we dealt with each cluster separately to compute the Jaccard coefficient similarity with all sentences within that blog post. Table 7.1 shows an example of a blog post consisting of eight sentences, while Table 7.2 shows its related generated clusters; then, Table 7.3 illustrates the Jaccard coefficient to measure the similarity between them.

Table 7.1. *Sentences of a blog post*

	A blog post	Length
1	Arcaptcha tqnyp AlkAbt$A b Allgp AlErbyp	5
2	mE bED Al<SrAr wAltEb w Aljd w Aljhd qAm Al$Abyn EbdAlEzyz Almqry w nASr Alwhyby mn jAmEp Albtrwl w AlmEAdn b<TlAq m$rwE Erby hw Al>wl mn nwEh fy EAlmnA	27
3	why EbArp tTwyr ltqnyp kAbAt$A mn xlAl dEmhA llgp AlErbyh w tmt tsmythA b Arcaptcha	14
4	hw mktbp brmjyp lmTwry AlmwAqE w AlAnZmp AlmtSlp bAl$bkp lltSdy llTlbAt Alwhmyp w AlAwtwmAtykyp	13
5	Hyv >n AlkAbt$A hw AxtbAr ymyz byn <jAbp Almstxdm Al>dmy w brAmj AlHAswb	12
6	lkn AlEdyd mnA >SAbh Alml1 mn AlHrwf AllAtynyp Alty knA nHn kErb mjbryn ElY AstxdAmhA kl mrp fAllgp hy wEA' HDArtnA	21
7	mn h*A AlmnTlq >tt fkrp Alm$rwE Aljdyd bhdf tkvyf AlmHtwY AlErby w tshyl <ySAlh ljmyE Almstxdmyn AlErb	17
8	>nA $xSyA >dEm h*h Alfkrp w >&ydhA b$dh	8

Table 7.2. *A collection of related clusters to one blog post*

	Collection of Related Clusters	Length
1	EbdAlEzyz Almqry	2
2	Arcptch tqnyp AlkAbt$A b Allgp AlErbyp	5
3	Al<SrAr wAltEb wAljd wAljhd	4
4	AlEdyd mnA >SAbh Alml1 mn AlHrwf AllAtynyp Alty knA nHn kErb mjbryn ElY AstxdAmhA kl mrp	16
5	nASr Alwhyby	2
6	>tt fkrp Alm$rwE Aljdyd bhdf tkvyf AlmHtwY AlErby w tshyl <ySAlh ljmyE Almstxdmyn AlErb	14
7	tTwyr ltqnyp kAbAt$A mn xlAl dEmhA llgp AlErbyh wtmt tsmythA b Arcptch	12
8	Al>wl mn nwEh fy EAlmnA	5
9	AxtbAr ymyz byn <jAbp Almstxdm Al>dmy w brAmj AlHAswb	8
10	>dEm h*h Alfkrp w >&ydhA b$dh	6
11	Al>wl mn nwEh tTwyr ltqnyp kAbAt$A	6
12	tqnyp AlkAbt$A	2
13	m$rwE Erby	2
14	mktbp brmjyp lmTwry AlmwAqE w AlAnZmp AlmtSlp bAl$bkp lltSdy llTlbAt Alwhmyp w AlAwtwmAtykyp	12

Table 7.3. *The Jaccard coefficient similarity between clusters (CL) and sentences (S)*

	Jaccard coefficient similarity between clusters (CL) and Sentences (S)													
S	CL 1	CL 2	CL 3	CL 4	CL 5	CL 6	CL 7	CL 8	CL 9	CL 10	CL 11	CL 12	CL 13	CL 14
1	0	1	0	0	0	0	1	0	0	0	0.5	0.5	0	0
2	0.1	0	0.11	0	0.1	0.03	0	0.2	0	0	0.11	0	0.14	0
3	0	0.32	0	0	0	0	0.86	0	0	0	0.2	0.13	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0.93
5	0	0.13	0	0	0	0	0.1	0	0.67	0	0.1	0.1	0	0
6	0	0	0	0.78	0	0.04	0.05	0	0	0	0	0	0	0
7	0	0.1	0	0.15	0	0.82	0	0.1	0	0.06	0	0	0.18	0
8	0	0	0	0	0	0.13	0	0	0	0.75	0	0	0	0

As can be seen from Table 7.3, results with a Jaccard coefficient similarity of 0.5 and above were considered. Thus, for this blog post we gathered 9 out 14 similar clusters from the original blog post.

Subsequently, we computed the accuracy of clustering for each blog post in our corpus. As represented in Choudhary and Bhattacharyya (2002), the accuracy of clustering measurement can be computed using the following formula:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{the\ number\ of\ clusters\ with\ high\ similarity\ scores}}{\boldsymbol{the\ number\ of\ all\ clusters}} \quad \textbf{(7. 5)}$$

The accuracy of the Jaccard coefficient shown in Figure 7.3 for the test dataset.

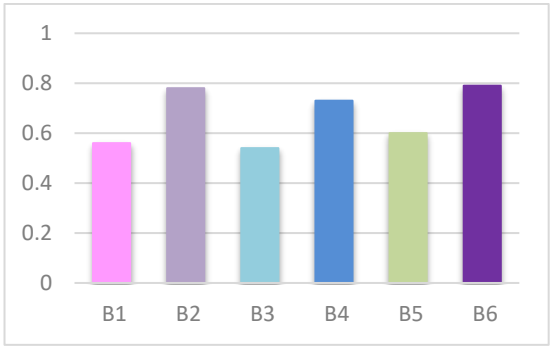


Figure 7. 3. *The accuracy of the Jaccard coefficient: the test dataset; (B) represents a blog post*

As a result, we found the range of the accuracy of the Jaccard coefficient similarity to be between 0.53 and 0.79 for our test dataset.

Cosine similarity

In order to measure the accuracy of clustering between a pair of items, a specific definition was required to measure the closeness between the pair in terms of either their 'distance' or its inverse, their 'similarity'. This measurement for the similarity or distance varied; measures such as the Cosine similarity have been applied widely (Salton, 1989; Strehl et al., 2000). In our case, we wanted to use the Cosine to measure the similarity between clusters and sentences within the blog post.

Before applying the Cosine similarity, it is important to mention the need for a document representation which refers to a way to model or represent a text document. For instance, it may be modelled as a 'bag of words', in which individual words are considered to stand alone, and where the order of these words is unnecessary (Baeza-Yates and Neto, 1999; Huang, 2008; Steinbach et al., 2000). In this bag of words, we count the words in the bag; thus, when compared to the definition of a set in the field of mathematics, a bag of words is different. Each word in the bag is considered a term, which matches a dimension in the data space results; and then, each document turns to a vector consisting of values; these should be non-negative on each dimension (Huang 2008).

In order to measure the weight of each term, we use term frequency. This means that the more frequently that a term appears, the more descriptive and important that term is in characterising what the document is about. This assumes that we have:

$D = \{d_1, d_2, \ldots. d_n\}$, where D is a number of documents.

$T = \{t_1, t_2, \ldots. t_n\}$, where T is the set of distinct terms appearing in D.

Hence, any document is represented as a dimensional-vector *(td)*, so, the frequency of a term ($t \in T$) in a document ($d \in D$) is denoted as *tf (d, t)*. Then, to represent a vector *(td)* of a document *(d)* as defined by Huang (2008):

$$\boldsymbol{td} = [\boldsymbol{tf}(\boldsymbol{d}, \boldsymbol{t1}), \boldsymbol{tf}(\boldsymbol{d}, \boldsymbol{t2}), \dots\dots, \boldsymbol{tf}(\boldsymbol{d}, \boldsymbol{tn})] \quad \textbf{(7. 6)}$$

When presenting documents as vectors, by this we compute the similarity degree of two documents regarding 'the correlation of their corresponding vectors; which can be further quantified as the Cosine of the angle between the two vectors' (Huang 2008).

As previously mentioned, 'terms' are 'words' but in reality, most terms that appear frequently are not necessarily the most important ones. Thus, the stop-words, such as *do*, *a*, *are*, *and*, and so on, which may appear more frequently than other words, need to be removed because they are unimportant (Baeza-Yates and Neto, 1999; Huang, 2008). Based on that, in order to determine and reflect the importance of these terms, rather than using the simple 'term frequency' of a term in a document, which is defined as [*tf (d, t)*], it is necessary to think instead of 'term frequency with inverse document frequency'; this calculates the weight of a term *t* in a document *d*. This is done using a factor that does not associate the importance of a term with the number of times it occurs in the whole document collection. A term that occurs often in a document is of interest, but if it also occurs in very many documents then it becomes of less interest (loses discriminative power); this latter property is captured by document frequency, expressed as inverse document frequency (i.e., rarer terms have higher inverse document frequencies). Combining term frequency and inverse document frequency yields the *tf.idf* measure. As defined by Huang (2008):

$$\boldsymbol{tf.idf} = \boldsymbol{tf}(\boldsymbol{d}, \boldsymbol{t}) \times \boldsymbol{log_{10}} \left(\frac{|\boldsymbol{D}|}{\boldsymbol{df}(\boldsymbol{t})} \right) \quad \textbf{(7. 7)}$$

where *df (t)* is the number of documents that contain the term (*t*). Thus, the *tf.idf* value is used instead of using the term frequency of each term separately in order to build the term vectors. This allows us to determine the importance of each term with respect to each document, according to its weight, i.e., its *tf.idf* value for that document.

Furthermore, we need to compute the mean value of all term vectors in the document. This can be done as described in Strehl et al. (2000) and Huang (2008) by normalising all vectors to a unified length in order to avoid domination of the long documents in the clusters. By considering *R* to be 'a number of documents', the length normalisation is defined as:

$$\overrightarrow{t_R} = \frac{1}{|R|} \sum_{\overrightarrow{t_d} \in R} \overrightarrow{t_d} \quad \textbf{(7. 8)}$$

Now, we have documents represented as 'term vectors', so it is time for the clustering process. The similarity between a cluster and an item—sentences within a blog post in our case—needs to be compared (Jain et al., 1999). This comparison can be made by using a popular similarity measurement, like the Cosine similarity, which can be applied to the text documents, and which has been widely used in the domain of IR applications and for clustering (Baeza-Yates and Neto, 1999; Larsen and Aone, 1999; Zhao and Karypis, 2002; Zhao and Karypis, 2004). The Cosine similarity can be computed as defined by Huang (2008):

$$Cosine\ (\overrightarrow{t_x}, \overrightarrow{t_y}) = \frac{\overrightarrow{t_x} . \overrightarrow{t_y}}{|\overrightarrow{t_x}| \times |\overrightarrow{t_y}|} \quad \textbf{(7. 9)}$$

where $\overrightarrow{t_x}, \overrightarrow{t_y}$, are two documents represented as 'dimensional vectors' for the set of terms: T = {t_1, t_2, …. t_n}. As Huang (2008) says, 'Each dimension represents a term with its weight in the document, which is non-negative. As a result, the Cosine similarity is non-negative and bounded between [0 and 1].'

In our research the Cosine was applied to measure the similarity between each cluster and the related sentences in the blog post. Assuming we had a cluster (x) and all related sentences (y), the approach we used to find the similarity between (x, y) was as follows:

1 We grouped each cluster with the related sentences only from the blog post in order to build a bag of words between them. By taking the example of a blog post in Table 7.1 and its collection of clusters in Table 7.2, we grouped them as shown in Figure 7.4. Our conditions for building this bag were: removing the stop-words[31], and considering any word that existed in (ASWN and TechTerms list) to appear in this bag, as well as selected unique words; an example is shown in Table 7.4.

2 After building this bag of words, each word in the bag was considered as a term; we then measured the similarity between a cluster (x) and all related sentences (y) by computing for both of them these formulas *tf.idf* (7.7), the *normalisation* (7.8) and then, by measuring the *Cosine* (7.9).

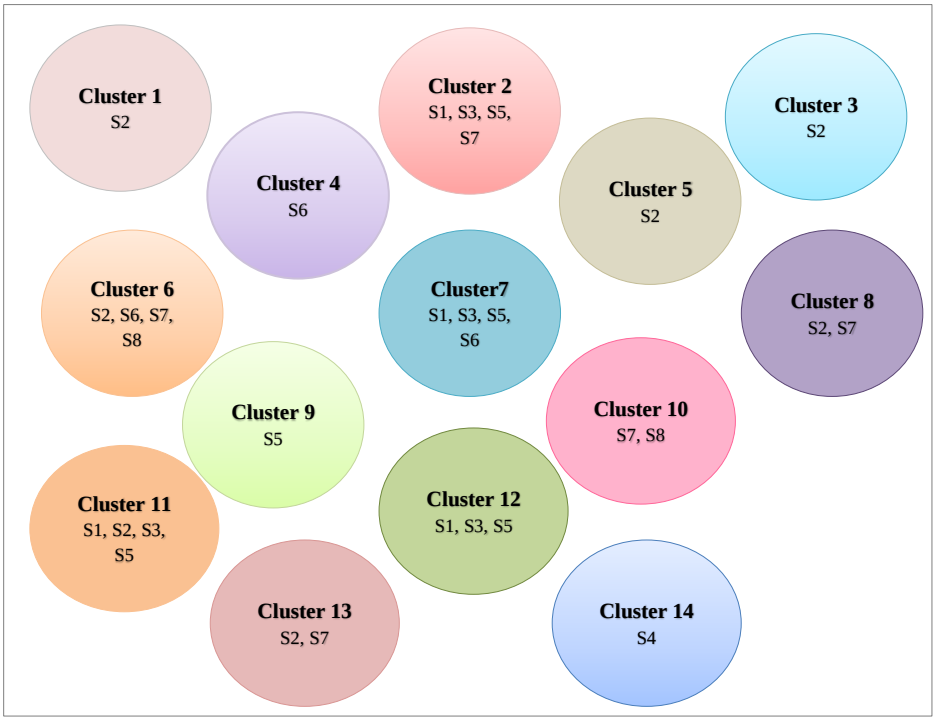


Figure 7. 4. *Grouping clusters with related sentences in a blog post to build a 'bag of words'*

[31] We used the Arabic stop-words list shown in Appendix C.

The determination of 'terms' from the bag of words can be explained through an example. Let us take, for example, Cluster 12 with the related sentences S1, S3, and S5.

Cluster 12: `tqnyp AlkAbt$A`

Sentences (1, 3, 5):

`Arcaptcha tqnyp AlkAbt$A b Allgp AlErbyp`
`why EbArp tTwyr ltqnyp kAbAt$A mn xlAl dEmhA llgp AlErbyh wtmt tsmythA b Arcaptcha`
`Hyv >n AlkAbt$A hw AxtbAr ymyz byn <jAbp Almstxdm Al>dmy w brAmj AlHAswb`

- The stop-words list was applied to remove words such as {`why, mn, xlAl, b, Hyv, >n, hw, byn, w`}.
- Words that existed in the ASWN and the TechTerms list were picked up, such as, {`AlkAbt$A, Arcaptcha, brAmj, AlHAswb, tqnyp`}.
- Words that were unique in the union of the sentences were also picked up, such as, {`Allgp, AlErbyp, tTwyr, AxtbAr, Almstxdm`}.

Now, after retrieving all the important 'terms', we computed the Cosine similarity as shown in Table 7.4.

Table 7.4. *Example of measuring the Cosine similarity between a cluster and related sentences*

Terms	Cluster 12			Sentences (S1,S3,S5)			Cosine Similarity
	tf	*tf.idf w*	*normalise*	*tf*	*tf.idf w*	*normalise*	
`AlkAbt$A`	1	0.9031	1	3	0.6291	0.33	0.13
`Arcaptcha`	0	0	0	2	0.7833	0.5	0
`brAmj`	0	0	0	1	0.9031	1	0
`AlHAswb`	0	0	0	1	0.9031	1	0
`Tqnyp`	1	0.9031	1	2	0.7833	0.5	0.4
`Allgp`	0	0	0	2	0.7833	0.5	0
`AlErbyp`	0	0	0	2	0.7833	0.5	0
`tTwyr`	0	0	0	1	0.9031	1	0
`AxtbAr`	0	0	0	1	0.9031	1	0
`Almstxdm`	0	0	0	1	0.9031	1	0
The Cosine similarity between Cluster 12 and sentences (S1,S3,S5) is :							*0.53*

We followed the same approach with the remaining 14 clusters, as shown in Figure 7.4. The results of the Cosine similarity between these clusters and related sentences for this particular blog post are shown in Figure 7.5.

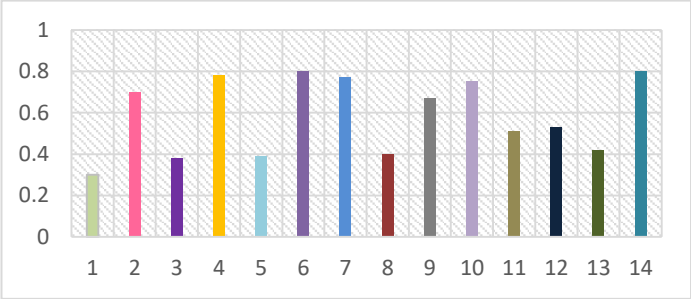


Figure 7.5. *Cosine similarity results between clusters and related sentences for a blog post (B)*

As can be seen, clusters with Cosine results that scored 0.5 and above were considered in the final cluster collection. The Cosine similarity showed us the 9 clusters (out of 14) that shared the most similarities to this blog post.

The remaining figures represent the Cosine similarity of the final collection of clusters, with similarity results of 0.5 and above, for each blog post from our test dataset, where (B) represent a blog post.

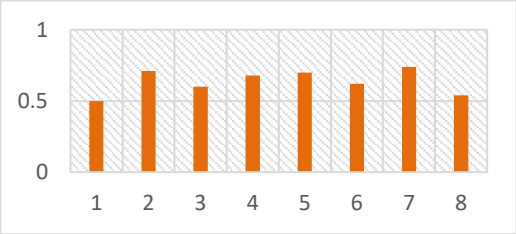


Figure 7. 6. *Cosine similarity results: test dataset (B1)*

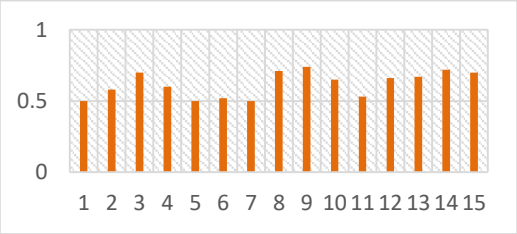


Figure 7. 7. *Cosine similarity results: test dataset (B2)*

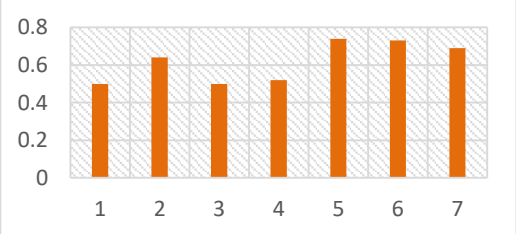


Figure 7. 8. *Cosine similarity results: test dataset (B3)*

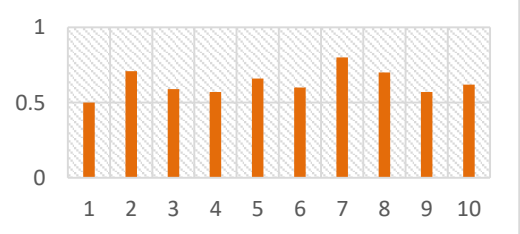


Figure 7. 9. *Cosine similarity results: test dataset (B4)*

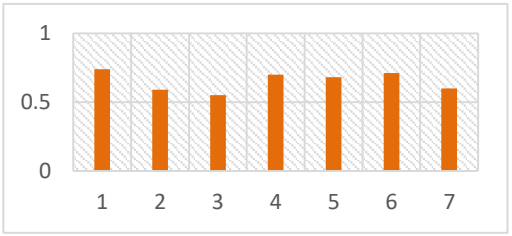


Figure 7. 10. *Cosine similarity results: test dataset (B5)*

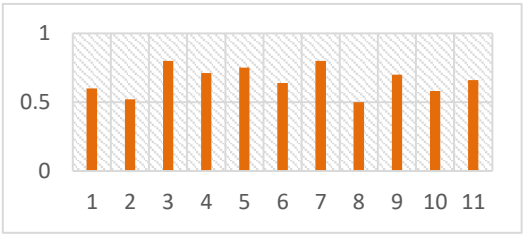


Figure 7. 11. *Cosine similarity results: test dataset (B6)*

As mentioned previously, for a long blog post the maximum number of clusters is 30, and for a short blog post the maximum is 14. From our test dataset of 6 blog posts, the following can be seen:

- The number of clusters obtained for the final collection of similar clusters ranged 7 to 15 clusters for the test dataset.
- The Cosine similarity results for the final collection of similar clusters ranged from 0.5 to 0.74 for the test dataset.

7.3 Evaluation

There are two types of clustering evaluations, internal and external (Färber et al., 2010; Manning et al., 2008; Pourrajabi et al., 2014). The internal evaluation is the function of clustering formalisation in order to obtain high similarity between an object and relative clusters to measure clustering quality (Manning et al., 2008; Pourrajabi et al., 2014). This was what we covered in the previous section; we evaluated the similarity between collections of clusters and original sentences within blog posts by using the Jaccard coefficient and the Cosine similarity to end up with a useful collection of clusters for each blog post.

In contrast, the external evaluation is the function of evaluating clusters with an external data that was not utilised for clustering i.e. a set of objects classified by human experts with a high level of inter-judgment agreement. This set, then, can be used as a

gold standard for evaluation. Then, we can determine how well the automatic clustering matches the gold standard (Manning et al., 2008; Pourrajabi et al., 2014). In our research, we built our gold standard for clustering by considering three main classes as shown in Figure 7.12.

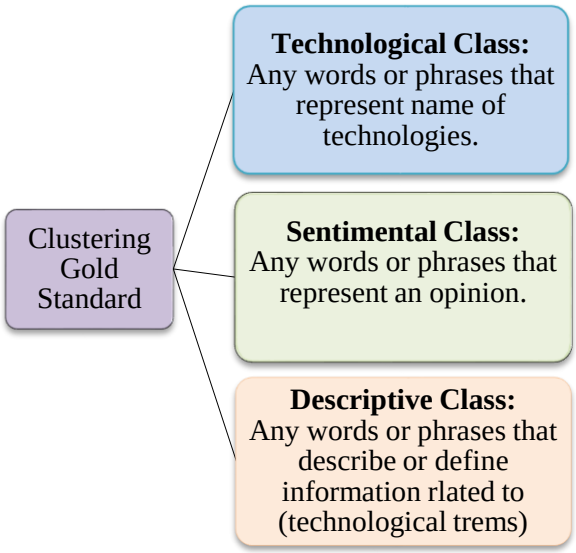


Figure 7. 12. *Clustering gold standard*

Our corpus was divided into two datasets, training and test, in which blog posts containing the same technological terms were used; we included most of them in the training dataset and some in the test dataset. For example, blog posts that included information about the iPhone—features, types, comparison with other types, prices, and so on—were placed primarily in the training dataset, with some reserved for the test dataset. Our datasets divided into 35 blog posts for training and 6 for the test. Thus, we aimed to include all different types of technological terms that appeared in our corpus in the training dataset, in order to build the clustering gold standard which could be also used for the test dataset.

Next, we used a group of three Arabic native speakers who were experts and qualified in the Arabic language to create human clusters from each blog post in the training dataset; these were based on the three main classes shown in Figure 7.12. Then, we computed the inter-judgment agreement for each class. The average Kappa score for

the training dataset for each class, together with the percentage of human clusters made for each class, is shown in Table 7.5.

Table 7.5. *Clustering gold standard evaluation*

	Technological Class	Sentimental Class	Descriptive Class
Human Clusters %	32.35 %	39%	28.65 %
Average Kappa Scores	0.884	0.755	0.817

The following can be seen:

- 32.35 percent of human clusters (group of words) belonged to the technological class, which means they represented names of technologies; the average inter-judgment agreement showed a Kappa score of 0.884.
- 39 percent of human clusters (group of words) belonged to the sentiment class with an average Kappa agreement of 0.755.
- 28.65 percent of human clusters (group of words) belonged to the descriptive class with a Kappa score of 0.817.

Next, the automatic clusters gathered from MALLET LDA topic modelling were evaluated against the clustering gold standard by taking the final collection of similar clusters for the test dataset which are shown in Figure 7.6 to Figure 7.11.

We used the following formulas to compute the F-measure (Manning et al., 2008):

$$\boldsymbol{F - measure = \frac{Precision * Recall * 2}{Precision + Recall}} \quad \textbf{(7. 10)}$$

Where:

$$\boldsymbol{Recall = \frac{\text{the number of correct clusters by MALLET}}{\text{the number of clusters made by annotators}}} \quad \textbf{(7. 11)}$$

$$\boldsymbol{Precision = \frac{\text{the number of correct clusters from MALLET}}{\text{the number of all clusters from MALLET}}} \quad \textbf{(7. 12)}$$

Tables 7.6 shows the average F-measure result for the test dataset.

Table 7.6. *The average result of F-measure for test dataset against human clusters*

The Average of automatic Clusters with Human Clusters		
Recall	*Precision*	*F-measure*
0.731	0.66	0.693

7.4 Discussion of the results

As can be seen, the average F-measure result for the automatic clustering against the human clusters was 0.693 for the test dataset, which we consider to be sufficiently high for the purposes of our rsearch.

It is worth mentioning that we also used the X-means algorithm in Weka[32] for clustering, but we found that the LDA topic modelling offerd by MALLET was providing us with our goal clusters (77 percent) better than the X-means; thus, the X-means was ignored and we continued our experiments with MALLET LDA topic modelling.

7.5 Summary

This chapter covered the second phase of our research, namely, the procedure of clustering, and explained how we ended up using a hybrid clustering method, one of our research contributions.

The overall procedure that we based our clustering phase on was the one offered by MALLET LDA topic modelling; this clustering extraction method was found to be suitable for our single-issue blogs and was able to offer a simplified way to analyse large amounts of unlabelled text.

[32] Weka, the *Waikato Environment for Knowledge Analysis*, is a popular open source machine learning toolkit (Witten et al., 2011).

Using MALLET LDA topic modelling, we began first by treating the raw text in our selected blogs as the first level of clustering. We followed this stage by another, in which only MWTs related to those blogs were treated; this was the second level of the clustering process. Finally, a combination of the raw text and related MWTs was treated, as the third level, in order to demonstrate the success of such integration in the case of Arabic MWTs with the raw text, and the ability to extract meaningful and useful clusters on the same theme.

After generating all clusters from our hybrid method, we applied a similarity filter, making use of the Jaccard coefficient and Cosine similarity so that we would be able to evaluate the final collection of clusters for the betterment of further processes in our framework. Then, an evaluation process was carried out to assess the similarity of this group of clusters against a gold standard of clusters made by our human experts. This procedure gave us 30 clusters for the long blog posts, and 14 for the short ones, a workable number from the high number of blogs, which ranged from 8 to 298 sentences.

Moreover, the hybrid method treatment provided us with the short strings in clusters that supplied us with name of technologies, for example. On the other hand, the long string of clusters provided us with definitions and/or opinions, as another example of the benefits of such arrangements of clusters. These results meant we were able to determine the technology that the blog was talking about, relevant information about it and any related opinion.

We progressed to the second cluster generation and evaluation stage, the aim of which was to assess and validate the quality of our results. To do this for the clusters themselves, we measured the quality of our gathered clusters, from each blog independently. To achieve this, we used the similarity filter process, using the Jaccard coefficient and the Cosine similarity. However, to assess the accuracy of the clustering process, itself, we utilised another research accuracy measuring formula, and then applied

it as a clustering measure for each blog against the total number of clusters. Form the Jaccard coefficient measure, we obtained an accuracy of the similarities ranging from 0.53 to 0.79 for our test dataset. Cosine similarity was used to measure the closeness between clusters and their corresponding sentences in a blog post. We obtained a Cosine similarity closeness score ranging from 0.5 to 0.74 for the test dataset.

What we covered up to this stage is an internal evaluation process, relating the clustering formalisation measurement to the quality of similarities with their corresponding sentences. The latter are treated as objects to their relevant clusters within each blog. On the other hand, if we did the same with an external object, for example clusters by humans, then this would be an external evaluation process. This would determine for us how well our automatic clustering matched the gold standard established by human experts. The average F-score that we obtained for the automatic clustering to that of the humans was 0.693 for the test dataset.

CHAPTER 8: ANALYSIS OF TWITTER DATA

In this chapter, we describe in detail Phase 3: *Analysis of Twitter Data* in our research framework, as shown in Figure 1.2. First, some background information will be provided. This will be followed by an explanation of how the MWTs were used as keyword search term inputs to the Twitter Application Programming Interfaces (APIs) to collect tweets related to each blog post. Throughout this phase, we made use of an Arabic converter to transliterate the text from Arabic script to a readable format of Arabic. Finally, the percentage of collected tweets regarding each blog post will be discussed.

8.1 Background: related work

Using the Internet, a vast number of individuals articulate themselves by means of different social media platforms. Since its introduction, Twitter has become progressively more popular among social networks. Users are able to write status messages called ‘tweets’ to publish little updates on their profiles, with a character restriction of 140 (Bollen et al., 2011; Mittal and Goel, 2011).

Within the Arab world, Arabs using Twitter generally use informal Arabic, which includes varieties of Arabic such as Gulf Arabic and Egyptian Arabic (Al-Sabbagh and Girju, 2012). Addressing dialectal Arabic results in more difficulties for researchers dealing with NLP, as these are dialects that are mostly spoken, they have no standardisation, they are composed using free text, and they demonstrate considerable variation from Modern Standard Arabic (MSA) (Zaidan and Callison-Burch, 2014).

Given this reality, and because of the variability and intricacy of sentiment indicators which may be contained within one tweet, sentiment evaluation for Twitter is not straightforward. Though brief, Twitter messages can include a considerable amount

of information in a compressed state (Bifet and Frank, 2010). Furthermore, tweets can also portray sarcasm, they may express a combination of polarities, or the sentiments they express may be vague (Refaee and Rieser, 2014; Sharma and Vyas, 2010).

Even though Arabic is regarded as one of the top ten languages[33] on the Internet in terms of use, it is viewed as comprising inadequate language content on the Web, unlike English (Elhawary and Elfeky, 2010) with a small number of Web pages which concentrate on Arabic reviews (Shoukry and Rafea, 2012).

To collect information through Twitter, Twitter's APIs were employed to acquire the necessary tweets; Twitter offers a search API which allows one to look for tweets in a particular language (Kumar et al., 2014). Setting Arabic as the language of choice allows the user to access Arabic tweets. In this case, it was essential for a large number of Arabic sentences to be acquired and for the classifier to be trained and capable of categorising any newly provided sentence. Twitter comprised one of the chief sources of acquiring a considerable amount of data (Shoukry and Rafea, 2012).

Table 8.1 shows a comparison between the top ten API's/Web-services[34] that can provide tweets.

Table 8.1. *Top 10 APIs/Web-services searching for tweets*

Web-Service	**Search Features**			
	Maximum age of tweets	***Advanced options***	***Filter the results***	***Free***
Topsy	Unlimited	Specify when, set key words, set the domain within which to search, specify tweets which include hyperlinks or pictures.	According to how long ago post was made.	√
ReSearch.ly	Several months	Based on key words in particular locations or communities.	According to sex, whether negative or positive, and/or most retweets.	×

[33] http://www.internetworldstats.com/
[34] http://freenuts.com/top-10-websites-to-search-old-tweets/

Tweepy	Unlimited	Real-time search can provide up-to-date tweets.	By using search queries to collect what is needed.	√
Google	Unlimited	Google Real-time search can provide a tweet database and '*site:Twitter.com'* can be used to discover old posts.	Adjusting the timeline, working from the top few updates, or reading the entire thread of conversations.	√
Yahoo	Unlimited	Real-time search provides a tweet database, while '*site:Twitter.com'* can be used to search for old posts.	By clicking a Twitter icon, search results are limited to tweets.	√
Bing	Unlimited	Including 'Twitter.com' in the search or looking for older tweets with '*site:Twitter.com'*.	Breaking down all of the 'Twitter.com' results according to nation of origin and language in which they are written.	√
Searchtastic	Several months	The search produces a maximum of 30 pages or 3,000 tweets.	Including all of a Twitter user's posts or those of all of the user's followers, all posts using a key word, or all posts from a specific user including a key word, with capacity to show URLs in full and search hashtags.	√
BackTweets	Several months	Searches for Twitter posts that include hyperlinks.	No fee or registration required to use the basic version, which produces posts from the previous fortnight. The subscription version allows an expanded search. This can be added to Favourites or Bookmarks.	×
Snap Bird	10 days	Allows search of followers' tweets, own direct messages and other users' favourites.	Twitter posts, direct messages and favourites.	√
FriendFeed	Unlimited	Search can be defined with the advanced search function.	Can be used without registration. Registration option allows saving, commenting, liking or sharing.	√

8.2 Twitter data: collection and analysis

To collect tweets we used the Twitter APIs Topsy and Tweepy, shown in table 8.1, which enabled us to gather a stream of 'real-time tweets' to retrieve the target tweets that would be relevant for our study. Four steps were followed:

1 Search queries were set up by using the MWTs related to each blog post as *keywords.* The MWTs extracted from ArTerMine were written in a transliteration format. To enable us to use them, we applied an Arabic converter[35], whose design we contributed to, to change the format from the Buckwalter format to readable Arabic and conversely from Arabic letters to the Buckwalter format. In order to measure the proportion of the tweets retrieved – that were in fact relevant to each blog post in our corpus – we measured the Precision based on the following formula:

$$\textbf{Precision} = \frac{Relevant\ tweets\ from\ the\ query}{Irrelevant\ tweets\ from\ the\ query} \quad \textbf{(8. 1)}$$

We detected that the Precision of tweets that were relevant to each blog post in our corpus was 0.329. It is important to mention that our Precision score for relevant tweets retrieval in Arabic using MWT query, is similar to the Precision scores – using MWT query – for relevant document retrieval in Arabic as shown in Ababneh et al. (2016), Boulaknadel (2008) and Boulaknadel et al. (2008b).

2 The gathered relevant tweets were then organised and grouped based on their blog posts (because in Phase 4, as we will explain in Chapter 9, we dealt with each blog post with its relevant tweets separately). The data comprised 35 groups of tweets for the training dataset and six groups for the test dataset.

3 Next, the collected tweets were checked manually and cleaned up in order to keep the content of tweets only and to eliminate irrelevant data such as @username, email

[35] The Java code for converting the Arabic transliteration is available in Appendix D.

addresses and URLs; the hash symbol (#) was also removed but we retained the word after the hash symbol; and finally to consider tweets that were formal. Examples are shown in Table 8.2.

Table 8.2. *Examples of the process of cleaning up the collected tweets*

Original Tweets	Goal Tweets	Irrelevant Data Elimination
@tarekyass1994: سامسونج تبدأ في ترقية جالكسي نوت 4 و اس 5 لنسخة http//:fb.me/4NYX9kcJz	سامسونج تبدأ في ترقية جالكسي نوت 4 و اس 5 لنسخة مارشميلو	@username, URL
@AJArabic: قريبا، رسائل #فيسبوك ماسنجر تتلاشى تلقائيا http//:aja.me/yujp	قريبا، رسائل فيسبوك ماسنجر تتلاشى تلقائيا	@username, URL, #
@ALTTAWHID: ماذا تعرف عن #قوقل #درايف؟ هو محرك أقراص ثابت يتبعك أينما تذهب، فهو يتيح لك الاحتفاظ بكل شيء ومشاركة أي شيء مع اصدقاءك. فهو مخزنك الشخصي.	ماذا تعرف عن قوقل درايف؟ هو محرك أقراص ثابت يتبعك أينما تذهب، فهو يتيح لك الاحتفاظ بكل شيء ومشاركة أي شيء مع اصدقاءك. فهو مخزنك الشخصي.	@username, URL, #
@iMokhles تجربة تطبيق مدونة تقنية ابل مراسلتنا عبر الايميل **mokhleshussien@aol.com**	تجربة تطبيق مدونة تقنية ابل	@username, email

As the use of dialectal/informal Arabic did not form part of this research, any informal language was converted into MSA. For example, the following tweet was written in informal language:

كفو! معد فيه حروف انقليزيه معفطه

We changed it to formal language, as follows:

ممتاز! لم تعد هناك حروف إنجلزية غير مفهومة

This tweet means in English (*Excellent! There were no longer any English letters that are incomprehensible*). The formality/informality of the language was important since our focus was on meaning (i.e. sentiments expressed).

4 Now that we had a group of tweets for each blog post, the tweets within each group were then classified manually into three categories: (1) positive, i.e. all tweets that expressed positive opinions; (2) negative, i.e. all those expressing negative opinions;

and (3) facts, i.e. those with neutral content. Tables 8.3 shows the number of tweets for each category for each blog post in our test dataset.

Table 8.3. *Categorisation of the test dataset with the number of tweets for each category*

Categorisation of the Test Dataset							
Blog post	*Positive Count*	*Negative Count*	*Neutral Count*	**Blog post**	*Positive Count*	*Negative Count*	*Neutral Count*
B36	9	4	3	**B39**	2	7	1
B37	6	5	0	**B40**	1	4	0
B38	0	0	9	**B41**	8	0	6

8.3 Discussion of the analysis

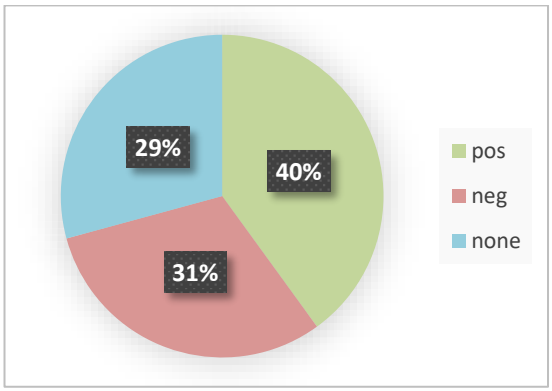


Figure 8.1. *Percentage of tweets by category: test dataset*

After categorising tweets into the three categories, it was found that for the test dataset, 40 percent of tweets were positive, 31 percent negative and 29 percent neutral. Results are shown in Figure 8.1.

Consistent with the literature, which reports that tweets are usually subjective – positive or negative (Mourad and Darwish, 2013; Shoukry and Rafea, 2012), in our research we obtained 71 percent in the test dataset.

8.4 Summary

In this chapter, we described in details Phase 3: *Analysis of Twitter Data*, in our research framework. As explained, the reason for choosing Twitter was that it has become increasingly popular among users of social networks, with users being able to write status messages called 'tweets' that contain whatever they wish to include.

As is common around the world, Arab users of Twitter use both formal and informal language. In the case of informal Arabic, this includes varieties of Arabic, such as Gulf Arabic and Egyptian Arabic. Addressing dialectal Arabic presents considerable difficulties, especially when dealing with NLP. This is because these dialects have no standardisation; they are composed using free-style text and they vary considerably from MSA. This meant that sentiment evaluation of Twitter data is an even more difficult task, with hidden perlocutionary acts that connote sentiments in non-MSA.

To collect information related to issues in blog posts from Twitter data, Application Programming Interfaces (APIs) were employed. This made it possible to gather a large number of Arabic sentences to be classified. Search queries were set up by using the MWTs related to each blog post as keywords. However, MWTs extracted from ArTerMine were written in a transliteration format, and to make use of them, we used an Arabic converter to change the transliterated text to Arabic script. The proportion of the tweets retrieved that were in fact relevant to each blog post was 0.329 (Precision). The gathered relevant tweets were then organised and grouped based on their blog posts. The data comprised 35 groups of tweets for the training datasets and 6 groups for test datasets. Next, the collected tweets were checked manually. Moreover, we eliminated irrelevant data such as @username, URLs, etc. and changed some of the informal tweets into formal MSA, for dialectal Arabic did not form part of this research.

Then the collected and refined tweets, for each blog post and within each group were classified into three categories: positive, expressing positive opinion; negative,

expressing negative ones; and finally facts, of neutral content. This resulted in the test dataset having 40 percent positive sentiments, 31 percent negative sentiments and 29 percent neutral. To compare our results with those in the published literature that confirmed the trend of subjectivity in Twitter data, our data showed the same trend, with 71% for the test dataset.

CHAPTER 9: LINKING, SENTIMENT CLASSIFICATION AND RANKING

This chapter describes in detail one of our research contributions (RC_1), which is addressed in the last phase – Phase 4 – of the research framework shown in Figure 1.2. The chapter presents some background information, followed by an overview of our experiment. The final collection of useful clusters for each blog post (the output of Phase 2) was taken, along with the related tweets collected for this particular blog post (the output of Phase 3); these were used as input for Phase 4. In other words, the analysis in this phase was applied to each blog post separately, whereby the Arabic SentiWordNet (ASWN) and the TechTerms list were applied at the lexicon look-up stage to determine words related to sentiments and to technology. Next, the collection of useful clusters and related tweets was applied at the next stage to classify sentiments in both. Finally, the sentiments were ranked, and the discussion of the final evaluation was carried out.

9.1 Background: related work

9.1.1 Linking news to social media based on content

The news provides most of the subjects of social media discourse (Balog et al., 2006; Java et al., 2007; Kwak et al., 2010; McLean, 2009; Phelan et al., 2009; Sayyadi et al., 2009; Thelwall, 2006). To a significant degree, even searches done on social media are the result of news occurrences and gossip items (Mishne and de Rijke, 2006). Consequently, the association between the two, social media and news, has begun to be the subject of close scrutiny from a variety of perspectives.

As has been illustrated by prior research, recent interest has concerned the forecasting of public reaction to news items published in social media (König et al., 2009; Szabó and Huberman, 2008; Tsagkias et al., 2009).

Research carried out by Tsagkias et al. (2011) looked at the links between news items and social media; they identified particular news items and looked for utterances on social media that made implicit reference to them. A three-step technique was applied in this study, which involved acquiring a number of query models from a particular source news article; these were subsequently employed in the extraction of utterances from a specified index of social media, bringing about multiple ranked catalogues that were subsequently merged, with the use of data fusion methods. Tsagkias et al. (2011) asserted that the formation of the query models was done by means of exploitation of the source article structure, and through employing overtly connected social utterances which specifically mentioned the source article. In an attempt to address query drift as a consequence of the considerable amount of text, whether from within the source news item itself or within media utterances connected to it directly, a graph-based technique was employed for the selection of discriminative terms. Tsagkias et al. (2011) generated query models with the use of data from *Twitter, Digg, Delicious, Wikipedia,* the *New York Times Community* and the blogosphere; this was done in an effort to illustrate that various query models comprising varying sources of data offered supplementary information, and that they were able to extract various social media utterances from the target index. The result is that techniques of data fusion are more capable of considerably enhancing retrieval performance than individual methods. Their experimental work is shown in Figure 9.1.

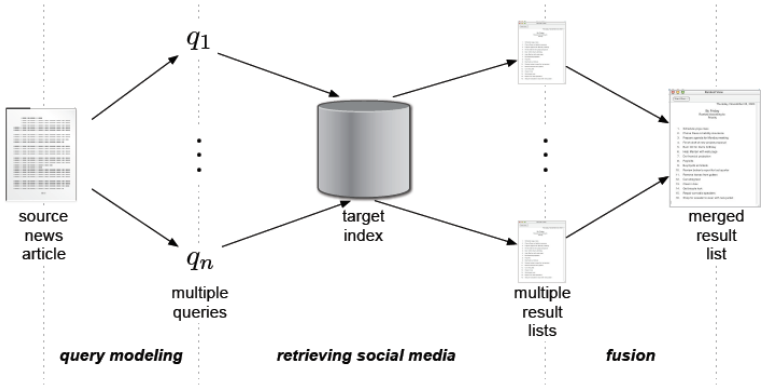


Figure 9.1. *Approach to finding linked social media utterances (Tsagkias et al., 2011)*

Ikeda et al. (2006) carried out an alternative study, in which they employed the similarity amid term vectors which signified blog posts and news items to resolve the availability of similarities connecting the two. In addition, Takama et al. (2006) employed the difference separating dates of publication for news articles and those of blog posts to formulate conclusions on the presence of a connection. A graph-based technique was employed by Gamon et al. (2008) to form a context for news articles drawn from blog posts. It has been demonstrated by other research that links recognition has been employed in the tracking of short cascades of information throughout the blogosphere (Adar et al., 2004; Gruhl et al., 2004; Kumar et al., 2004; Leskovec et al., 2007).

Regarding studies in the Arabic language, Elsawy et al. (2014) built TweetMogaz, which comprises a live tweets news portal and is available to the public at (http://www.tweetmogaz.com). It gathers tweets in Arabic and produces reports on news events occurring in various Arabic-speaking areas (e.g. Syria, the UAE and Egypt) as well as on international sports. Every day, a stream of as many as 12 million Arabic tweets are amassed, and online processing is carried out to filter information to various topics, producing inclusive reports on every topic and recognising contentious stories on every subject. The website is completely in Arabic, as it is targeted at the Arabic-speaking area.

Automatic updating is done on all of the website's information at 15-minute intervals to stay abreast of the Twitter trends (Elsawy et al., 2014).

In our research, we were interested in investigating the existence of the link between social networks themselves by using two types of social networks (blogs and Twitter). This was done by finding and identifying content within Twitter that reflected what the blog posts were about, and the opinions about it. Our research also differs from the previous studies, as we considered the sentiment analysis field to measure the sentiments in both (blogs and Twitter). In normal cases, people have to log into Twitter to get to know how people would want particular issues to be addressed in blogs. There is, therefore, a need for a new way of accessing information other than going to Twitter and searching on your own to find out about the opinions of others in blogs. Our framework was able to provide that information.

9.1.2 Sentiment classification, sentiment lexicon and classification sentiment in Arabic social media

Sentiment classification

Sentiment classification comprises an opinion-mining operation related to the identification of the general sentiment orientation (SO), if any, of the views encompassed in a particular document. Overall, the supposition is made that the document, when subjected to scrutiny, is found to contain subjective data, like that identified in feedback forms or product evaluations (Ohana and Tierney, 2009). Opinion orientation may be classified as making up part of contradictory negative or positive polarities—negative or positive criticism concerning a product, or complimentary or uncomplimentary views regarding a subject—or graded in line with a broad range of likely views, for instance on film reviews, with the responses from viewers varying between one and five stars (El-Beltagy and Ali, 2013; Ohana and Tierney, 2009).

Two key methods have been documented in literature when contemplating the job of establishing SO. The primary method employs a sentiment lexicon in addition to a part-of-speech (POS) tagger to establish if the direction of a provided text is negative, positive or neutral (objective). Normally, the sentiment lexicon is comprised of a catalogue of opinionated words; these are categorised as positive, negative or objective terms. The next technique is subject to the accessibility of text tagged with SO, which is subsequently employed in the training of a classifier to establish the SO of new and unseen text inside a similar domain to the tagged corpus (Matsumoto et al., 2005; Taboada et al., 2011).

The effort observed in Pang et al. (2002) signifies a number of supervised learning algorithms employing 'bag of words' representations familiar in research on text mining, with superior performance acquired by the employment of SVM, applied with unigrams as features. Investigation has also been carried out on how part-of-speech tagging (POS) could be exploited to improve sentiment classification. In Wilson et al. (2005), POS information is employed as an element of a set of features for carrying out sentiment classification on a data set of articles retrieved from newswire services. The researchers used comparable techniques to those employed in Salvetti et al. (2004), Kennedy and Inkpen (2006) and Gamon (2004) on various datasets. Turney (2002) used a method which traces and scores POS sequences to acquire elements for sentiment classification, with a comparable notion implemented in extraction of opinions for features of products as observed in Dave et al. (2003), Pang and Lee (2004), Abbasi et al. (2008a) and Yang et al. (2007).

Liu (2015) provides a comprehensive appraisal of sentiment evaluation study. He describes the challenge of sentiment evaluation as encompassing related sentiment terms like sentence subjectivity, opinion, opinion holder, object, emotions, etc. Additionally, he discusses the more widespread two-phase sentiment and subjectivity classification

technique at varying granularities (document and sentence extents) employing various machine-learning techniques (supervised and unsupervised) in addition to various means of building the necessary data resources (corpora and lexicon) (Liu, 2015).

Additionally, discovering sentiment may be formally described as discovering the quadruple {s;g;h;t} (Liu, 2012), in which **s** denotes the sentiment, **g** denotes the target object for which the sentiment is articulated, h denotes the holder (i.e the one articulating sentiment) and **t** signifies the time during which the sentiment was articulated. The target may comprise an entity, like the general topic of the review, or an aspect of an entity, which could comprise any feature or aspect of that entity. The resolution is formed subject to the application domain available. For instance, in product reviews, the product alone is normally the entity, while all things associated with that product (e.g. price, quality, etc.) are elements of the said product (Liu, 2012; Liu, 2015).

Aspect-level sentiment analysis is related not only to discovering the overall sentiment related to an entity, but also to discovering the sentiment for aspects of that entity. Therefore, aspect-based sentiment analysis is increasing in popularity as it results in very fine-grained sentiment information which could have practical applications in various fields (Schouten and Frasincar, 2016). Overall, three processing stages may be differentiated when carrying out aspect-level sentiment evaluation: identification, classification and aggregation (Tsytsarau and Palpanas, 2012). Tsytsarau and Palpanas (2012) asserted that in practical terms, not all methods apply all of the three stages or in this precise sequence. They show the key points of aspect-level sentiment analysis:

1 The initial stage is associated with the identification of sentiment–target pairs within the text.

2 The subsequent stage comprises the classification of the sentiment–target pairs. The articulated sentiment is classified in line with a predetermined group of sentiment

values, for example negative and positive. At times the target is categorised in line with a predetermined group of aspects too.

3 The sentiment values are aggregated for every aspect to offer a precise overview. The real presentation is reliant on the application's particular needs and requirements.

Schouten and Frasincar (2016) carried out a survey which concentrated on aspect-level sentiment analysis, in which the objective is the discovery and aggregation of sentiment on entities stated in documents or aspects of the documents. Schouten and Frasincar (2016) offer recent solutions that are classified according to whether they offer a technique for aspect detection, sentiment evaluation or a combination of the two.

This was not a current concern for our research because we had just focused on 'sentiment classification'. However, we propose looking into these aspects in the future, as discussed in Chapter 10. However, it depends (for Arabic) on having not only good NER but also event extraction, which is not yet well-researched for Arabic (Aliane et al., 2013). We note also that those attempts that do exist appear to concentrate on detection of events in the sense of disasters or attacks, rather than on the linguistic notion of event. Nevertheless, Baradaran (2013) addressed information extraction in historical texts and also, the Automated Content Extraction conference (ACE) had an activity on Arabic event extraction (Linguistic Data Consortium, 2005a). However, Maybury (2010) reports low F-scores for Arabic relation extraction (and no activity to speak of for event extraction), based on ACE results.

Sentiment lexicons

Sentiment lexicons comprise resources which relate sentiment orientation (SO) with words. The use of sentiment lexicons in opinion mining stems from the conjecture that singular words may be viewed as a component of opinion information, and thus may offer indications regarding document sentiment and subjectivity.

Opinion lexicons that are manually formed tend nonetheless to be limited to a smaller number of terms. Through its very nature, construction of manual lists is a long-term endeavour, and there is the potential for annotator prejudice (Pang et al., 2002). In order to address such matters, lexical induction techniques have been suggested within the literature; these have the objective of extending the magnitude of opinion lexicons from a basic group of seed terms, either by means of investigating term relationships, or through appraising similarities within document corpora (Kennedy and Inkpen, 2006). Research work within this area, as observed in Hatzivassiloglou and McKeown (1997), provides a catalogue of constructive and unconstructive adjectives by means of the assessment of conjunctive statements within a document corpus.

A further technique in widespread use comprises the acquisition of opinion terms from the WordNet database of relationships and terms (Miller et al., 1990), generally through the examination of the semantic associations of terms like synonyms and antonyms. Lexicons constructed employing this method may be viewed as being implemented for subjectivity detection study in Wilson et al. (2005) and employed for sentiment classification in Dave et al. (2003) and Salvetti et al. (2004).

As observed by Rao and Ravichandran (2009), one result of term associations within the WordNet database is a greatly disconnected graph; therefore, extension of opinion information from a core of seed words through the scrutiny of semantic associations like synonyms and antonyms is inclined to be only limited to a subset of terms. So as to address this challenge, information encompassed within term glosses—explanatory text following every term—may be investigated to imply term direction, subject to the supposition that a provided term and the terms comprised within its gloss are inclined to signify similar polarity. A technique for lexicon expansion is suggested in Andreevskaya and Bergler (2006), in which terms are allocated negative or positive opinions subject to the availability of terms known to carry opinion substance available

in the term gloss. The writers assert that glosses entail a potentially reduced extent of noise, as they are planned to match as nearly as feasible the elements of significance of the word, and they have comparatively standard grammar, syntactic arrangement and style. This concept has also been observed in Esuli and Sebastiani (2005), this time through the employment of supervised learning techniques for extending a lexicon through the investigation of gloss information, providing constructive accuracy enhancements above a gold standard. This is a comparable technique to that utilised in the construction of the SentiWordNet opinion lexicon (Esuli and Sebastiani, 2006). SentiWordNet was constructed through a two-stage method; primarily, WordNet term associations like synonym, hyponymy and antonym are investigated to extend a core of seed words employed in Turney and Littman (2003), and known *a priori* to contain constructive or unconstructive opinion prejudice. Following a set number of iterations, a subset of WordNet terms is acquired with either a constructive or unconstructive label. The glosses of these terms are subsequently employed in the training of a committee of machine-learning classifiers. To reduce subjectivity, the classifiers are trained with the employment of various algorithms and various training set sizes. The forecasts from the classifier committee are subsequently employed to establish the SO of the rest of the terms within WordNet (Esuli and Sebastiani, 2006).

Sentiment classification in Arabic social media

A considerable amount of effort has been expended on sentiment classification intended for Arabic social media (Hammad and Al-awadi, 2016; Salameh et al., 2015; Taboada, 2016). Studies have been carried out by Abdul-Mageed et al. (2014), who trained an SVM classifier using a manually labelled dataset and implemented two-phase classification, which initially isolates subjective from objective sentences and subsequently classifies the subjective into constructive or unconstructive instances. A number of datasets have

been collected by the writers from several social media resources; these include tweets, chatroom messages, Wikipedia Talk pages and forum posts. Nonetheless, these resources have not been made available to the public as yet.

An additional study carried out by Mourad and Darwish (2013) trained Naïve Bayes and SVM classifiers using Arabic tweets annotated by two native Arabic speakers. Tweets were manually annotated by Refaee and Rieser (2014) for sentiment with the use of two Arabic native speaking individuals. An SVM was employed to classify tweets employing a two-stage technique, polar vs. neutral, which was followed by positive vs. negative. Abbasi et al. (2008a) concentrated on carrying out sentiment classification at document extent. They employed 56 syntactic, stylistic and morphological features of Arabic to carry out the classification. A sentence-extent classification was carried out by Abdul-Mageed et al. (2011) for MSA. Their conclusion was that the emergence of a positive or negative adjective, based on their lexicon, comprises the most significant feature.

In alternative endeavours, Abdul-Mageed et al. (2012) broadened their work to social text. Their conclusions were that: (1) regarding sentiment classification, POS tags are not as efficient as in subjectivity classification; and (2) the majority of dialectal Arabic tweets are unconstructive. Finally, they anticipated that broadening/conforming polarity lexicons to novel domains, such as social media, would bring about considerably higher gains.

Kok and Brockett (2010) initiated a random-walk-base method to produce paraphrases from parallel corpora. These were observed to be more efficient in the production of additional paraphrases through the traversing of conduits comprising lengths longer than two. El-Kahky et al. (2011) implemented graph fortification (McGlohon et al., 2011) on the challenge of transliteration mining to infer mappings which were not observed in training.

In our research, in order to classify sentiment, after applying our sentiment lexicon (the ASWN), we classified sentiment using the NB classifier.

9.1.3 Measuring sentiment strength

Within the literature associated with sentiment analysis, a distinction is made between two kinds of textual analyses: those in which a supposition is made that the text signifies a viewpoint and thus only requires computation of the robustness of its polarity (Thelwall et al., 2010; Thelwall et al., 2012), and those within which prior to the quantification of polarity, it has to be established whether the text is subjective or objective (Montejo-Raez et al., 2014; Thelwall, 2013). A broad overview of research regarding sentiment evaluation is available in Pang and Lee (2008), Liu (2010) and Tsytsarau and Palpanas (2012).

The majority of the suggested uses of polarity classification determine a level of negativity or positivity. In some cases, a degree of neutrality is also generated (Duwairi and Alshboul, 2015; Montejo-Raez et al., 2014).

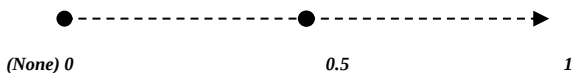
In our research, it was determined that for the quantification to be used in the calculation of the robustness of sentiment, an actual value in the interval [0, 1] would be adequate. Values nearer to one would mirror a strong positive or negative sentiment articulated within the text. However the nearer to a zero value a post is, the more neutral it would be (Duwairi et al., 2015). Therefore, in our case, tweets that were related to every blog post were ranked based on the degree of sentiments; as a ranking scale for sentiment evaluation was necessary, we employed a formula comparable to that used by Moreno-Ortiz and Hernández (2013) and Zhu et al. (2013), taking into consideration the sentiment values for every word within the ASWN, as illustrated in this formula:

$$\boldsymbol{Ranking\ Scale} = \sum_{\boldsymbol{w \in S}} \frac{\boldsymbol{Number\ (w) * (Sentiment\ Scores)}}{\boldsymbol{Length\ (S)}} \quad \textbf{(9. 1)}$$

Where

w = sentiment words, S = Sentence

Thus, the sentiment scale range was as follows:



9.2 Method

Our proposed method for this phase followed four steps:

1 ***The linking process.*** We had already gathered the final collection of useful clusters for each blog post in our corpus in Phase 2; we had also retrieved tweets related to blog posts in Phase 3. It was now time to pass them to Phase 4, which worked per blog post by taking its related useful clusters (Group 1: each cluster considered as a sentence) with related tweets (Group 2: each tweet treated as a sentence) as input for this phase; this is shown in Figure 1.2, Phase 4.

2 ***The use of lexicons.*** In this step two lexicons were applied to our method:

(1) The TechTerms list. This list was used as a lexicon look-up, in order to pick up all terms related to technology from final useful clusters related to blog posts; these were later evaluated using F-measure, and the results were compared to those obtained from the manual annotation task.

(2) The ASWN. This was used as a sentiment lexicon to determine all sentiment words in our datasets from final useful clusters related to blog posts and tweets; these were later classified using the NB classifier.

3 ***The sentiment classification process.*** In this step we classified sentiments for each blog post (its useful clusters, and related tweets). We trained the NB classifier (Manning et al., 2008; Manning et al., 2009; Shimodaira, 2015) as a second step, after applying the ASWN, as follows:

(1) We trained the NB classifier on sentences from Group 1 (useful clusters) and Group 2 (related tweets).

(2) The sentences in each group were separated into objective and subjective opinions.

(3) The subjective sentences then classified into positive and negative opinions.

(4) The average F-measure was calculated for both groups: useful clusters and tweets.

4 ***Ranking of sentiment process.*** In this step, we already had the sentiment classified in tweets based on subjectivity (positive or negative) and objectivity. Next, formula 9.1 was used to rank tweets based on the average degree of positivity, the average degree of negativity and the average degree of neutrality, as shown in Figure 9.2.

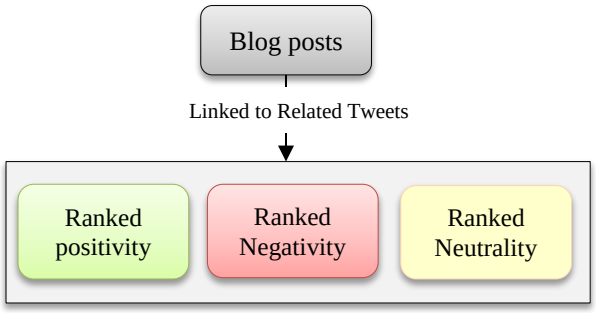


Figure 9.2. *Ranking tweets based on sentiments*

9.3 Evaluation and results

This section presents a discussion of the results obtained from evaluating technological terms, sentiment classification and ranking tweets. These were then compared to the manual annotation results achieved by experts.

9.3.1 Technological terms: results and discussion

The results obtained for technological terms after we applied the TechTerms list to blog posts on the test dataset are shown in Table 9.1 and Figure 9.3.

Table 9.1. *Technological terms results*

	Technological Terms	
	Test data	*Humans*
Recall	0.91	***Kappa Score:*** 0.868
Precision	0.7998	
F-measure	0.851	

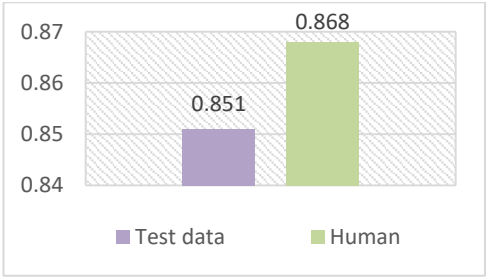


Figure 9.3. *Technological terms results*

As can be seen, the F-score obtained from our corpus for the test dataset was 0.851. When comparing these results to those obtained from the manual annotation task for technological terms, which had a Kappa score of 0.868, we noticed the results obtained were what we had aimed to achieve from building the TechTerms list. Table 9.2 shows

the percentage of technological terms types (class types shown in Chapter 5 in Figure 5.4) that appeared in our corpus.

Table 9.2. *The percentage of technological terms in the research corpus*

English Technologies	English Companies	Arabic Technologies	Arabic Companies
In English:	*In English:*	13%	6%
34%	12%		
Transliterated to Arabic:	*Transliterated to Arabic:*		
26%	9%		

As shown in Table 9.2, most technological terms that appeared in our corpus, both in the training and the test datasets, were from the *English technological* category; of the total number, 34 percent were technological terms written in English and 26 percent were transliterated into Arabic. The second group of terms that appeared was from the *English companies* category; 12 percent were written in English and 9 percent were transliterated. In contrast, technological terms in Arabic accounted for 13 percent of technological terms, and Arabic companies for 6 percent. From this, we can conclude that most technological terms written in Arabic technology blogs were represented either in the English language or transliterated to Arabic.

It is important to mention that the TechTerms list was applied as a lexicon so as to determine and pick up any technological terms only from our corpus—blog posts—and not from tweets. The reasons for that are: (1) to compare the technological terms result obtained from applying the TechTerms list to blog posts, with the manual annotation result performed by experts; and (2) for the related tweets gathered, some of them did not include the name of this technology, but had opinions or facts related to this technology which we were more concerned about from tweets.

9.3.2 Sentiment: results and discussion

We evaluated sentiments using the NB classifier in our corpus and in related tweets.

The corpus

The sentiment results obtained for our corpus for the test dataset are shown in Table 9.3 and Figure 9.4.

Table 9.3. *Sentiment results for the corpus: test dataset*

	Sentiment Type	Recall	Precision	F-measure	Subjectivity Average
Objective	*Neutral*	0.831	0.8112	0.821	
Subjective	*Positive*	0.837	0.775	0.805	0.799
	Negative	0.793	0.796	0.794	

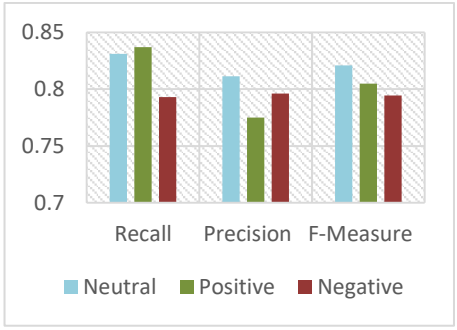


Figure 9.4. *Sentiment results for the corpus: test dataset*

As can be seen, we achieved our projected result for sentiment for the test dataset as the average sentiment result for subjectivity was an F-score of 0.799. The objectivity F-score was 0.821 for the test dataset. When these results were compared to the Kappa scores of the expert annotation results, shown in Table 9.4 and Figure 9.5, we noticed that the use of a sentiment lexicon we built, the ASWN, has contributed to improve the sentiment results from our experiment. This improvement of the results answered our *RQ2* and confirmed our hypothesis *RH2*. This means that improving the world knowledge of resources and tools that are available for Arabic sentiment analysis has brought about an

improvement in the results obtained. Thus, there is a direct relation between tools and resources and improving the average sentiment results.

Table 9. 4. *Comparison of the sentiment results with experts annotation results*

	Comparison of Sentiment Results	
	Human Evaluation *(Kappa Score)*	**Test Data** *(F-measure)*
Objectivity	0.789	0.821
Subjectivity	0.776	0.799

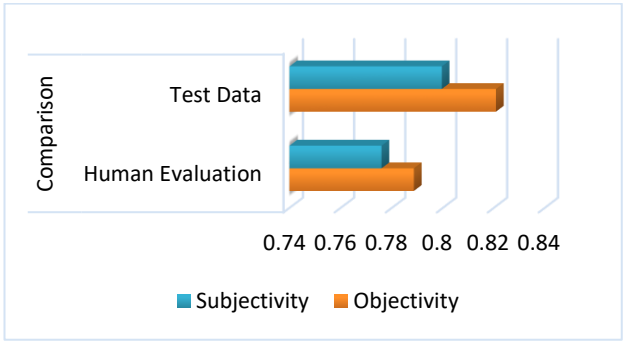


Figure 9. 5. *Comparison of the sentiment results with experts annotation results*

Tweets

Figure 9.6 shows the sentiment results obtained for related tweets, for the test dataset.

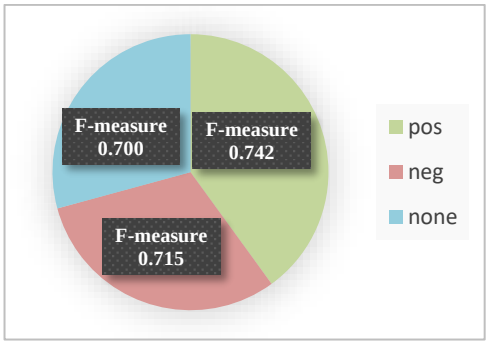


Figure 9.6. Sentiment results for tweets: test dataset

As can be seen, for the test dataset, we found before the percentage of tweets categories in section 8.3 Chapter 8 that 40 percent of positive tweets had a sentiment result with an F-measure of 0.742; the F-measure results for the 31 percent of negative tweets and the 29 percent of neutral tweets were 0.715 and 0.700 respectively.

9.3.3 Ranking: results and discussion

The last step of the evaluation for Phase 4 of our research framework was to rank, for each blog post, the related tweets based on the degree of sentiment. As discussed earlier (section 9.1.3), formula 9.1 was used; the ranked results for tweets related to each blog post (B) for the test dataset are shown in Table 9.5 and Figure 9.7.

Table 9.5. *Ranking of tweets based on the degree of positivity and negativity: test dataset*

Ranking of Tweets Based on the Degree of Positivity and Negativity – Test dataset					
Blog post	*Rank Positivity*	*Rank Negativity*	**Blog post**	*Rank Positivity*	*Rank Negativity*
B36	0.5625	0.25	**B39**	0.2	0.7
B37	0.545	0.456	**B40**	0.143	0.571
B38	0	0	**B41**	0.583	0

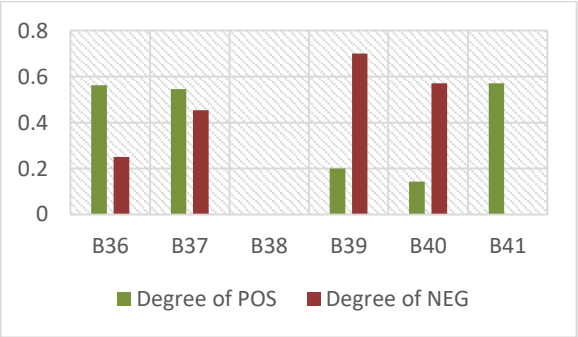


Figure 9.7. *Ranking of tweets based on the degree of positivity and negativity: test dataset*

Results were as follows:

- 3 out of 6 were ranked in the range of 0.54 – 0.58 for positivity; these are considered positives.
- 2 out of 6 were ranked in the range of 0.57 – 0.7 for negativity; these are considered strong negatives.

9.4 Summary

To sum up, this chapter has discussed the fourth phase of our research framework which was divided into four steps: (1) linking each blog post with related tweets; (2) applying the TechTerms list and the ASWN lexicon; (3) classifying sentiments in both blog posts and tweets; and (4) ranking related tweets based on the degree of sentiments.

The technological terms results obtained in terms of F-measure from our corpus for the test dataset was 0.851. When compared to those obtained from the manual annotation task for technological terms, which had a Kappa score of 0.868, from these results we met the goal that we had aimed to achieve by building a high quality TechTerms list.

In addition, we achieved our desired results for sentiments for the test dataset as the average sentiment F-score for subjectivity was 0.799. The objectivity F-score was 0.821 for the test dataset. Comparing these results to the Kappa scores of the expert annotation, we noticed that the development of the ASWN, a sentiment lexicon, provided us with the sentiment results that we wanted to achieve, which answered our RQ_2 and confirmed our hypothesis RH_2. The implication is that improving global awareness of resources and tools that are available for Arabic sentiment analysis can improve the results obtained. Thus, there is a direct relation between the availability and use of tools and resources and improvements in sentiment results.

We noticed that we obtained more tweets classed as positive (i.e., between 0.54 and 0.58) for the test dataset. The remaining tweets were classed as negative (in the range from 0.57 to 0.7) for the test dataset.

CHAPTER 10: CONCLUSION AND FUTURE WORK

This research aims to shed light on the field of opinion mining and sentiment analysis and to investigate links between Arabic blogs and Twitter, based on the subject or matter that discussed in the blog post and opinions expressed towards this discussed issue from Twitter. This chapter provides a summary of the thesis, which is followed by validation of the research hypotheses and research questions. Finally, suggestions for future work are made.

10.1 Thesis summary

In this thesis, we focused on the domain of opinion mining and sentiment analysis in order to investigate two forms of Arabic social media: blogs and Twitter. Our studies were aimed at linking the content from blog posts to related tweets that discussed the same information and expressed opinions about it. Three fundamentals aspects were covered comprehensively in this thesis: (1) the research corpus; (2) the tools and resources required for this research; and (3) the research framework.

Through this research we contributed first to the development of a corpus based on Arabic technology blogs. The corpus is based on formal MSA Arabic and comprises 2,350 sentences from 41 blog posts in Arabic. As stated by Alhazmi and McNaught (2013) five well-known technology Arabic blogs were employed in the creation of our corpus; the corpus was divided into two datasets: 35 for training and six for test purposes. Arabic technology blogs were selected as the focus of the research because they presented a number of challenges. The first was the use of English within the blogs; to address this, we constructed an open list which included names of technology, company names and names of software. Next, since blog writers employed transliteration for a significant

proportion of foreign company names and technology terms, transliterations were included within this list. Where more than one transliteration for a term existed, each had to be indexed and included. A final challenge was misspelling; the data was checked by professionals in the Arabic language.

A team of three annotators was employed to assemble and annotate our corpus. An increased level of inter-annotator concurrence was made available through the establishment of the TechTerms task, particularly in the marking of technology and company names (written in English, transliterated into Arabic or written in Arabic). Nonetheless, discrepancies were observed in differentiating between sentiment tasks and facts. The Kappa statistic was employed in the evaluation of the corpus by humans. The outcomes were documented and reported. Results showed that overall there was substantial agreement among the annotators regarding the facts and sentiments tasks, and an almost perfect agreement within the TechTerms task. The achieved Kappa score of 0.868 is considered satisfactory. Nonetheless, the outcomes from the facts and sentiments task were anticipated to be within a similar range of Kappa scores, while the outcomes acquired comprised 0.789 and 0.776 correspondingly for facts and sentiments.

The second fundamental aspect was the tools and resources used in this research. We contributed through this research to an improvement in some of the tools and resources for Arabic sentiment analysis: (1) a sentiment lexicon, the ASWN (Alhazmi et al., 2013); (2) the Arabic multi-words-terms extraction the ArTerMine; and (3) the technological terms list TechTerms.

The third aspect was the research framework, which consisted of four phases. The first phase was the initial process that our corpus went through. We started with the raw Arabic texts which had to be prepared by removing XML tags and correcting spelling mistakes manually. Then, we used the Arabic POS tagger via the U-Compare workflow. Finally, the ArTerMine was applied to extract MWTs for each blog post.

The second phase was a hybrid clustering method, which is also considered to be one of our research contributions; this phase included two stages:

1 Clustering. This was carried out by employing MALLET LDA topic modelling and through integrating and combining three levels of clustering: (1) at the first level, employing the raw text from a blog post merely as the initial level to yield clusters of single words; (2) at the second level, employing MWTs associated with this subject to yield clusters of MWTs; and (3) at the third level, implementing the two simultaneously to yield clusters of MWTs. Using the result of this hybrid technique, we verified the adequacy of combining Arabic MWTs with the objective of acquiring both significant and practical clusters which mirror the subject matter of a blog post.

2 Similarity. This was employed to evaluate the final entire collection of clusters from our hybrid technique, as we employed the Jaccard coefficient as well as Cosine similarity. Here, the reason for employing similarity was to acquire a set of practical useful clusters.

In order to evaluate the final outcome of useful clusters, we asked native speakers to create reference clusters – we refer to these as 'human clusters'. From the human clusters, we noticed that (1) 32.35 percent of human clusters made up part of the technological class, signifying that they portrayed names of technologies; the average inter-judgment agreement comprised a Kappa score of 0.884; (2) 39 percent of human clusters formed part of the sentiment class, with the average Kappa agreement comprising 0.755; and (3) 28.65 percent of human clusters made up the descriptive class, with a Kappa score of 0.817.

Then, the the automatic clusters gathered from MALLET LDA topic modelling were evaluated against the clustering gold standard (human clusters). The average outcomes for F-measure of the automatic clustering against the human clusters comprised 0.693 for the tests.

The third phase of our research framework was the collection and analysis of tweets; this was done using the Twitter APIs, which facilitate the collection of a stream of 'real-time' tweets, in order to acquire target tweets pertinent to our research subject. The following stages were implemented:

1 Search queries were established through employment of the MWTs associated with every blog post as comprising keywords. The MWTs retrieved from the ArTerMine were composed in a transliteration arrangement; therefore, so as to be able to employ them, we used an Arabic convertor to change the text to comprehensible Arabic. To evaluate this query, we found out that the Precision of tweets that were relevant to each blog post in our corpus was 0.329.

2 Subsequently, the collected tweets were arranged and grouped in relation to their blog posts. There were 35 divisions of tweets for the training dataset and six for the test dataset.

3 Then, the tweets that had been gathered were appraised manually and were cleaned up individually so as to only maintain the substance of tweets and to delete irrelevant information such as @username, the hash (#) symbol, URLs, and email addresses. The remaining words were assessed to ensure that all contained only official MSA.

4 Tweets in dialectal Arabic were excluded. These were not analysed during this study as they comprise informal language; the language was altered into MSA.

Subsequently, we had a group of tweets for every blog post, and the tweets in this division were then divided into three classes: (1) positive, i.e. all tweets comprising favourable opinions; (2) negative; i.e. those expressing unfavourable opinions; (3) facts (neutral), i.e. those comprised of content that is factual rather than opinion-based. Following the categorisation of tweets within the three classes, we were able to establish that within the test dataset, 40 percent comprised positive tweets, 31 were negative and 29 percent were neutral.

The fourth phase is considered to be one of our research contributions; this was divided into four steps: (1) linking each blog post with related tweets; (2) applying the TechTerms list and the ASWN lexicon; (3) classifying sentiments in both blog posts and tweets; and (4) ranking related tweets based on the degree of sentiments.

The technological terms outcomes acquired for F-score from our corpus for the test dataset comprised 0.851. This compares with the outcome acquired using the manual annotation task for technological terms, which had a Kappa score of 0.868. From the outcomes obtained by using the TechTerms list, we met our aim which we intended to address by constructing a high quality TechTerms list.

Furthermore, we addressed our intended outcomes for sentiments on the test dataset as comprising the average sentiment outcome for subjectivity which yielded F-score 0.799 for the test. However, the objectivity outcome showed F-score of 0.821 for the test dataset.

By reference to the expert annotation, it was observed that the inclusion of a specially constructed sentiment lexicon, the ASWN, considerably enhanced the sentiment outcomes from our test. This enhancement of the outcomes responded to our RQ_2 and substantiated our hypothesis RH_2, which signifies that enhancing world awareness of instruments and resources accessible for Arabic sentiment evaluation has enhanced the average sentiment outcomes.

We observed that we received more tweets that were positive; these ranked from 0.54 – 0.58 for the test dataset. However, the remainder of the tweets were ranked as negatives and were within the range of 0.57 – 0.7 for the test dataset.

10.2 Confirmation of research hypotheses

Our research hypotheses were confirmed, as shown in this section.

> ***RH1 :***
>
> It should be feasible to link content in different types of social media such as blogs and micro-blogs by using text mining techniques, using the Arabic language as a case study.

Confirmation of *RH1*. We confirmed that by using text mining techniques, we were able to establish a link between the content from Arabic blog posts with related information from tweets. Specifically:

1 From the use of the ArTerMine tool, the MWTs extracted from each blog post were used for retrieving tweets from Twitter (micro-blogs) that reflect the same issue in the blog post. By using this tool, we collected all related tweets that were needed to establish the link between blogs posts and Twitter.

2 Through implementing Phase 4 of our research framework, we were able to formalise the datasets from blogs posts and Twitter within one platform.

3 Text mining techniques also helped in classifying the sentiments in both social media (blogs and Twitter) using the ASWN and the NB classifier.

Based on these facts, with the use of text mining techniques we were able to link different types of social media such as blogs and micro-blogs. Hence, our first research hypothesis has been confirmed.

RH2 :

Analysis of implicitly expressed emotions can improve sentiment-based techniques in linking different types of social media, using the Arabic language as a case study.

Confirmation of *RH2*. We confirmed our hypothesis through exploring the lack of sentiment resources that were available for the Arabic language and then building the ASWN sentiment lexicon to be used and examined in this research. Thus, we noticed the following:

1 We met our projected results for sentiment for the test dataset as the average F-score for subjectivity was 0.799; and for objectivity was 0.821.

2 When comparing the Kappa score for the expert annotation, we noticed that the use of the ASWN improved the sentiment results.

Thus, this improvements of the results has confirmed hypothesis *RH2*, which means improving world knowledge of resources and tools available for Arabic sentiment analysis has improved the results obtained. Thus, there is a direct relation between tools and resources and improving the average sentiment results.

RH3 :

Involving multi-words-terms in a hybrid clustering method should enhance the quality of the outcome clusters.

Confirmation of *RH3*. We confirmed this hypothesis through the use of the ArTerMine tool by using the MWTs related to each blog post in the process of building clusters. Thus, we have noticed the following:

1 The use of MWTs provides us with useful and short clusters, i.e. the name of the technology that the blog post contains.

2 Also, the use of the raw text within a blog post, and then with its MWTs, provides us with some long clusters that include, for example, definitions of this technology or opinions about it.

Thus, by the hybrid clustering method for each blog post in our corpus we identified the following: (1) the technology discussed in the blog post; (2) related information regarding this technology; and (3) opinions about this technology.

So, from this, we can confirm that our generated clusters for each blog post represent and determine information about this particular blog post content; this suggests that involving the MWTs with the hybrid clustering method helped in improving the quality of the outcome clusters with our projected information.

In confirming our research hypotheses, we answered research questions *RQ1* and *RQ2*: by enhancing the techniques used for text mining, we identified and classified sentiments expressed in Twitter which reflected the content and opinions in blog posts.

10.3 Future work

Our future directions could concentrate on the following:

1 Expanding our research scope to other languages, such as English. This could be done by replacing the tools we used for Arabic with English tools, such as an English POS tagger, SentiWordNet and the original English version of the TerMine tool.

2 Expansion of the use of social media networks beyond blog posts and Twitter. For example, we could establish links between blog content and Facebook by applying the same techniques we used for Twitter.

3 Classification of sentiment using the SVM classifier. The time constraints for this research allowed only the use of the NB classifier for classifying sentiments although it provided us with our projected results.

4 Experimentation with our proposed framework on different domains, such as politics; we anticipate helpful results with this particular domain.

5 Achieving wider coverage regarding our sentiment lexicon, the ASWN, by considering the SLSA (Eskander and Rambow, 2015) entries to expand our lexicon. However, the licence that Eskander and Rambow (2015) offer, while it allows building upon the SLSA, which would allow integration in the ASWN, disallows redistribution, which means we could not distribute an ASWN that incorporated the SLSA. However, as we disagree with the way in which the sentiments were evaluated in the SLSA, we would prefer to involve experts to evaluate SLSA entries.

6 Introduction of a majority class for our research. The outcome of our hybrid clustering method was evaluated against a gold standard (the expert humans clusters) due to the lack of a majority class (Kelaiaia1 and Merouani, 2016; Manning et al., 2008; Pourrajabi et al., 2014). The ASWN and the TechTerms list could be expanded to build a majority class that would enable the efficient use of other external criteria of

clustering quality, such as purity and entropy (Färber et al., 2010; Manning et al., 2008; Pourrajabi et al., 2014).

7 Replacement of remaining manual steps with automatic processing to enable an end-to-end system. This was not the focus of our research which concentrated on proof of concept and on core elements of automatic processing to investigate our research hypotheses. There is also a large topic to be tackled by the field of Arabic NLP in relation to robust processing of raw, informal text containing orthographic and other errors. Investigation of aspect-based sentiment analysis would determine how this might improve linking of tweets and blogs. This, however, would involve a step-change in Arabic NLP to deliver event extraction of sufficient quality, thus this is likely to be a longer term goal of future work.

Through working on this thesis, we gained a lot of experience and knowledge in the domain of opinion mining and sentiment analysis, i.e. discovering its various techniques, encountering the different challenges, and learning how to think practically to find solutions to these challenges among others. More precisely, we made the difficult choice of working on the Arabic language. Our prior knowledge that there was a gap and lack of the resources and tools available for Arabic put us in the position of having to contribute some tools and resources that we hope to be useful and used practically in this regard. Furthermore, we hope this research will prove to be of benefit to the field.

APPENDIX A: AN EXAMPLE OF THE ARTERMINE OUTPUT FOR ONE BLOGPOST

The reals are C-value scores indicating significance of the MWT within the text.

2.07944	fkrp Alm$rwE Aljdyd hdf tkvyf AlmHtwY AlErby
1.94591	jAmEp Albtrwl AlmEAdn <TlAq m$rwE Erby
1.88334	mktbp brmjyp mTwry AlmwAqE Al>nZmp AlmtSlp Al$bkp AltSdy
1.84839	mktbp brmjyp mTwry AlmwAqE Al>nZmp AlmtSlp Al$bkp
1.84207	mktbp brmjyp mTwry AlmwAqE Al>nZmp AlmtSlp Al$bkp AltSdy AlTlbAt
1.79176	Al$Abyn EbdAlEzyz AlmqrY nASr Alwhyby
1.79176	tshyl <ySAl jmyE Almstxdmyn AlErb
1.79176	<jAbp Almstxdm Al|dmy brAmj AlHAswb
1.79176	bED Al<SrAr AltEb Aljd Aljhd
1.76901	mktbp brmjyp mTwry AlmwAqE Al>nZmp AlmtSlp
1.70267	AlmwAqE Al>nZmp AlmtSlp Al$bkp AltSdy AlTlbAt
1.66792	Al>nZmp AlmtSlp Al$bkp AltSdy AlTlbAt Alwhmyp
1.66355	hw mktbp brmjyp mTwry AlmwAqE Al>nZmp AlmtSlp
1.66355	AlmwAqE Al>nZmp AlmtSlp Al$bkp AltSdy AlTlbAt Alwhmyp
1.65393	mktbp brmjyp mTwry AlmwAqE Al>nZmp
1.64792	hw mktbp brmjyp mTwry AlmwAqE Al>nZmp AlmtSlp Al$bkp
1.62887	AlmwAqE Al>nZmp AlmtSlp Al$bkp AltSdy
1.62887	Al>nZmp AlmtSlp Al$bkp AltSdy AlTlbAt
1.62159	hw mktbp brmjyp mTwry AlmwAqE Al>nZmp
1.60944	EbArp tTwyr tqnyp kAbAt$A
1.60944	xlAl dEmhA Allgp AlErbyp
1.5986	mktbp brmjyp mTwry AlmwAqE Al>nZmp AlmtSlp Al$bkp AltSdy AlTlbAt Alwhmyp
1.53579	hw mktbp brmjyp mTwry AlmwAqE
1.53506	hw mktbp brmjyp mTwry AlmwAqE Al>nZmp AlmtSlp Al$bkp AltSdy
1.50214	Al>nZmp AlmtSlp Al$bkp AltSdy
1.50214	mktbp brmjyp mTwry AlmwAqE
1.49448	AlmwAqE Al>nZmp AlmtSlp Al$bkp
1.43061	Al$bkp AltSdy AlTlbAt Alwhmyp
1.40826	hw mktbp brmjyp mTwry
1.38629	arcaptcha
1.38629	AstxdAmhA kl mrp
1.38629	Al>nZmp AlmtSlp Al$bkp AltSdy AlTlbAt Alwhmyp Al>wtwmAtykyp
1.34382	Al$bkp AltSdy AlTlbAt Alwhmyp Al>wtwmAtykyp
1.31333	Al>nZmp AlmtSlp Al$bkp
1.30475	mktbp brmjyp mTwry
1.30475	AlmwAqE Al>nZmp AlmtSlp
1.28755	AltSdy AlTlbAt Alwhmyp Al>wtwmAtykyp
1.28755	Alm$rwE Aljdyd hdf tkvyf
1.28727	Al$bkp AltSdy AlTlbAt
1.26027	AltSdy AlTlbAt Alwhmyp
1.23226	hw mktbp brmjyp
1.19895	hw mktbp brmjyp mTwry AlmwAqE Al>nZmp AlmtSlp Al$bkp AltSdy AlTlbAt
1.19451	Alm$rwE Aljdyd hdf tkvyf AlmHtwY
1.18825	Alm$rwE Aljdyd hdf
1.15525	AlTlbAt Alwhmyp Al>wtwmAtykyp
1.10904	Albtrwl AlmEAdn <TlAq
1.10904	hdf tkvyf AlmHtwY
1.10904	AlmEAdn <TlAq m$rwE
1.09861	hw mtwfr
1.09861	AlEdyd mnA
1.09861	AlmwAqE Al>nZmp AlmtSlp Al$bkp AltSdy AlTlbAt Alwhmyp Al>wtwmAtykyp
1.09861	wEA' HDArp
1.09861	xlAl AlmtSfH
1.09861	AlHrwf AllAtynyp
1.09861	AltHmyl Al{TlAE
1.09861	Erb mjbryn
1.09861	tjrbp AlbrnAmj
1.07296	Albtrwl AlmEAdn <TlAq m$rwE
1.07296	fkrp Alm$rwE Aljdyd hdf
1.05085	Al>nZmp AlmtSlp
1.04368	AlmwAqE Al>nZmp
1.04079	mktbp brmjyp
1.04079	Al$bkp AltSdy
1.03972	fkrp Alm$rwE Aljdyd
1.03399	AltSdy AlTlbAt
1.0141	AlTlbAt Alwhmyp
0.98875	hw mktbp
0.97654	Alm$rwE Aljdyd
0.96129	AlmEAdn <TlAq
0.96129	hdf tkvyf
0.94167	Albtrwl AlmEAdn
0.94167	tkvyf AlmHtwY
0.94167	<TlAq m$rwE
0.9242	Al<SrAr AltEb Aljd
0.9242	<ySAl jmyE Almstxdmyn

0.9242	tkvyf AlmHtwY AlErby
0.9242	EbdAlEzyz AlmqrY nASr
0.9242	<TlAq m$rwE Erby
0.9242	jAmEp Albtrwl AlmEAdn
0.89588	fkrp Alm$rwE Aljdyd hdf tkvyf
0.87889	jmyE Almstxdmyn
0.87889	fkrp Alm$rwE
0.87889	AlmqrY nASr
0.87889	EbdAlEzyz AlmqrY
0.87889	Al<SrAr AltEb
0.87889	AltEb Aljd
0.87889	<ySAl jmyE
0.82396	jAmEp Albtrwl
0.82396	m$rwE Erby
0.82396	AlmHtwY AlErby
0.80472	hdf tkvyf AlmHtwY AlErby
0.80472	jAmEp Albtrwl AlmEAdn <TlAq
0.80472	AlmEAdn <TlAq m$rwE Erby
0.73241	bED Al<SrAr
0.73241	dEmhA Allgp
0.73241	Aljd Aljhd
0.73241	tTwyr tqnyp
0.73241	Al$Abyn EbdAlEzyz
0.73241	tshyl <ySAl
0.73241	Almstxdmyn AlErb
0.73241	<jAbp Almstxdm
0.73241	nASr Alwhyby
0.73241	Allgp AlErbyp
0.69315	Al$Abyn EbdAlEzyz AlmqrY
0.69315	qAm
0.69315	AltEb Aljd Aljhd
0.69315	>SAb
0.69315	>&yd
0.69315	ymyz
0.69315	>dEm
0.69315	knnA
0.69315	nwE
0.69315	tmt
0.69315	>tt
0.69315	bED Al<SrAr AltEb
0.69315	jmyE Almstxdmyn AlErb
0.69315	tshyl <ySAl jmyE
0.69315	<jAbp Almstxdm Al|dmy
0.69315	AlmqrY nASr Alwhyby
0.54931	tqnyp kAbAt$A
0.54931	xlAl dEmhA
0.54931	EbArp tTwyr
0	Alm$rwE Aljdyd hdf tkvyf AlmHtwY AlErby
0	kl mrp
0	Al<SrAr AltEb Aljd Aljhd
0	AlkAbt$A Allgp
0	hw mktbp brmjyp mTwry AlmwAqE Al>nZmp AlmtSlp Al$bkp AltSdy AlTlbAt Alwhmyp
0	jAmEp Albtrwl AlmEAdn <TlAq m$rwE
0	brAmj AlHAswb
0	dEmhA Allgp AlErbyp
0	fkrp Alm$rwE Aljdyd hdf tkvyf AlmHtwY
0	tshyl <ySAl jmyE Almstxdmyn
0	Albtrwl AlmEAdn <TlAq m$rwE Erby
0	<jAbp Almstxdm Al|dmy brAmj
0	EbArp tTwyr tqnyp
0	AlkAbt$A Allgp AlErbyp
0	xlAl dEmhA Allgp
0	<ySAl jmyE Almstxdmyn AlErb
0	mktbp brmjyp mTwry AlmwAqE Al>nZmp AlmtSlp Al$bkp AltSdy AlTlbAt Alwhmyp Al>wtwmAtykyp
0	EbdAlEzyz AlmqrY nASr Alwhyby
0	AstxdAmhA kl
0	tTwyr tqnyp kAbAt$A
0	bED Al<SrAr AltEb Aljd
0	Al$Abyn EbdAlEzyz AlmqrY nASr

APPENDIX B: SOME OF EXISTING LISTS INCLUDE TECHNOLOGICAL TERMS/COMPANIES IN (ENGLISH OR ARABIC) LANGUAGES

English:

- http://fortune.com/2015/06/13/fortune-500-tech/
- http://www.nesta.org.uk/blog/top-100-tech-companies?gclid=CKyW1snH184CFaMW0wodBCkFjg
- http://www.growthbusiness.co.uk/the-entrepreneur/business-leaders/2388638/25of-the-most-exciting-technology-companies-in-the-uk.thtml
- https://en.wikipedia.org/wiki/Category:Technology_companies
- https://en.wikipedia.org/wiki/Category:Computer_companies
- https://en.wikipedia.org/wiki/Category:Election_technology_companies
- https://en.wikipedia.org/wiki/Category:Information_technology_companies
- https://en.wikipedia.org/wiki/Category:Mobile_technology_companies
- https://en.wikipedia.org/wiki/Category:Software_companies
- https://en.wikipedia.org/wiki/Mobile_app

Arabic:

- http://arabhardware.net
- http://ar.itp.net
- http://www.tech-wd.com/wd/
- http://www.unlimit-tech.com/
- https://ar.wikipedia.org/wiki/قائمة_أكبر_شركات_التقنية_العالمية
- http://www.almrsal.com/post/301900
- https://ar.wikipedia.org/wiki/تصنيف:شركات_تقنية
- http://arabia-it.com/
- http://www.v22v.net/cat-9-1.html
- http://www.arabtech-group.com/arabic/portal/home.aspx

APPENDIX C: ARABIC STOP-WORD LIST

Arabic Stop-Words List						
fY	wkAnt	AmA	AlAwlY	mA	,	AElnt
Fy	llAmm	Ams	byn	mE	E$r	bsbb
Kl	fyh	AlsAbq	*lk	msA'	Edd	HtY
Lm	klm	AltY	dwn	h*A	Edp	A*A
Ln	lkn	Alty	Hwl	wAHd	E$rp	AHd
Lh	wfy	Akvr	Hyn	wADAf	Edm	Avr
Mn	wqf	AyAr	Alf	wADAft	EAm	brs
Hw	wlm	AyDA	AlY	fAn	EAmA	bAsm
Hy	wmn	vlAvp	Anh	Qbl	En	gdA
Qwp	whw	Al*Aty	Awl	qAl	End	$xSA
kmA	why	AlAxyrp	Dmn	kAn	EndmA	SbAH
lhA	ywm	AlvAny	AnhA	ldY	ElY	ATAr
mn*	fyhA	AlvAnyp	jmyE	nHw	Elyh	ArbEp
Wqd	mnhA	Al*Y	AlmADy	h*h	ElyhA	AxrY
wlA	mlyAr	Al*y	Alwqt	wAn	zyArp	bAn
nfsh	lwkAlp	AlAn	Almqbl	wAkd	snp	Ajl
lqA'	ykwn	AmAm	Alywm	kAnt	snwAt	gyr
mqAbl	ymkn	AyAm	f	wAwDH	tm	b$kl
hnAk	mlywn	xlAl	w	mAyw	Dd	HAlyA
wqAl	Hyv	HwAlY	w	B	bEd	bn
wkAn	Akd	Al*yn	qd	A	bED	bh
nhAyp	AlA	AlAwl	lA	>	AEAdp	vm
Af	An	Aw	Ay	bhA	Sfr	>n

APPENDIX D: THE ARABIC TRANSLITERATION CONVERTER

Convert from (Arabic letters to Buckwalter string)

```
package org.awn.transliteration;

import java.util.HashMap;
import java.util.Map;

import javax.swing.JOptionPane;

public class ArBwMap{
	private static boolean SHOW_UNTRANSLITERATED_CHARS = false;
	private static String SUBSTITUTE_CHAR = "?";

	static boolean usingDiacritics;


	private static Map<String, String> theMap = new HashMap<String,
String>();
	static {
		theMap.put(" "," ");
		theMap.put("\u0621", new String("'"));
		theMap.put("\u0622", new String("|"));
		theMap.put("\u0623", new String(">"));
		theMap.put("\u0624", new String("&"));
		theMap.put("\u0625", new String("<"));
		theMap.put("\u0626", new String("}"));
		theMap.put("\u0627", new String("A"));
		theMap.put("\u0628", new String("b"));
		theMap.put("\u0629", new String("p"));
		theMap.put("\u062A", new String("t"));
		theMap.put("\u062B", new String("v"));
		theMap.put("\u062C", new String("j"));
		theMap.put("\u062D", new String("H"));
		theMap.put("\u062E", new String("x"));
		theMap.put("\u062F", new String("d"));
		theMap.put("\u0630", new String("*"));
		theMap.put("\u0631", new String("r"));
		theMap.put("\u0632", new String("z"));
		theMap.put("\u0633", new String("s"));
		theMap.put("\u0634", new String("$"));
		theMap.put("\u0635", new String("S"));
		theMap.put("\u0636", new String("D"));
		theMap.put("\u0637", new String("T"));
		theMap.put("\u0638", new String("Z"));
		theMap.put("\u0639", new String("E"));
		theMap.put("\u063A", new String("g"));
		theMap.put("\u0640", new String("_"));
		theMap.put("\u0641", new String("f"));
		theMap.put("\u0642", new String("q"));
		theMap.put("\u0643", new String("k"));
		theMap.put("\u0644", new String("l"));
		theMap.put("\u0645", new String("m"));
		theMap.put("\u0646", new String("n"));
		theMap.put("\u0647", new String("h"));
		theMap.put("\u0648", new String("w"));
		theMap.put("\u0649", new String("Y"));
```

```
		theMap.put("\u064A", new String("y"));
		theMap.put("\u064B", new String("F"));
		theMap.put("\u064C", new String("N"));
		theMap.put("\u064D", new String("K"));

		theMap.put("\u067E", new String("P"));
		theMap.put("\u0686", new String("J"));
		theMap.put("\u06A4", new String("V"));
		theMap.put("\u06AF", new String("G"));

		theMap.put("\u0670", new String("`"));
		theMap.put("\u0671", new String("{"));
		// Numbers
		theMap.put("0", "0");
		theMap.put("1", "1");
		theMap.put("2", "2");
		theMap.put("3", "3");
		theMap.put("4", "4");
		theMap.put("5", "5");
		theMap.put("6", "6");
		theMap.put("7", "7");
		theMap.put("8", "8");
		theMap.put("9", "9");

		theMap.put("\u0660", "0");
		theMap.put("\u0661", "1");
		theMap.put("\u0662", "2");
		theMap.put("\u0663", "3");
		theMap.put("\u0664", "4");
		theMap.put("\u0665", "5");
		theMap.put("\u0666", "6");
		theMap.put("\u0667", "7");
		theMap.put("\u0668", "8");
		theMap.put("\u0669", "9");


		setUsingDiacritics(true);
	}

	/**
	 * Adds/removes short vowel letters to the HashMap and their
	 * respective Buckwalter equivalents
	 * @param using - (Boolean) True if using diacritics; false if not
	 */
	public static void setUsingDiacritics(boolean using) {
		usingDiacritics = using;
		if(usingDiacritics) {
			theMap.put("\u064E", new String("a"));
			theMap.put("\u064F", new String("u"));
			theMap.put("\u0650", new String("i"));
			theMap.put("\u0651", new String("~"));
			theMap.put("\u0652", new String("o"));
			theMap.put("\u064B", new String("F"));
			theMap.put("\u064C", new String("N"));
			theMap.put("\u064D", new String("K"));
		}
		else {
			theMap.remove("\u064E");
			theMap.remove("\u064F");
			theMap.remove("\u0650");
			theMap.remove("\u0651");
```

```
                    theMap.remove("\u0652");
                    theMap.remove("\u064B");
                    theMap.remove("\u064C");
                    theMap.remove("\u064D");
              }
       }

       /**
        * Transliterates an Arabic script string to Buckwalter by
referencing the HashMap.
        * @param s - (String) The Arabic script string to transliterate
        * @return result - (String) The Buckwalter string
        */
       public static String transliterate(String s) {
              String ret = "";
              for (int i=0;i<s.length();i++) {
                    String nextCh =
theMap.get(Character.toString(s.charAt(i)));
                    if (nextCh==null) {
                           if (SHOW_UNTRANSLITERATED_CHARS) {
                                  nextCh = Character.toString(s.charAt(i));
                           } else {
                                  nextCh = SUBSTITUTE_CHAR;
                           }
                    }
                    ret +=nextCh;
              }
              return ret;
       }

       public static void main(String[] argv) {
              String inputBW = JOptionPane.showInputDialog("Enter a string
in Unicode Arabic script");
              JOptionPane.showMessageDialog(null,
BwArMap.transliterate(inputBW));
       }
}
```

Convert from (Buckwalter string to Arabic letters)

```
public class BwArMap {
       private static boolean SHOW_UNTRANSLITERATED_CHARS = false;
       private static String SUBSTITUTE_CHAR = "?";
       private static boolean USING_LATIN_DIGITS = false;
       private static boolean USING_DIACRITICS = true;
       private static Map<String, String> theMap = new HashMap<String,
String>();
  static {
    theMap.put(" "," ");
       theMap.put("'","\u0621");
       theMap.put("|","\u0622");
       theMap.put(">","\u0623");
       theMap.put("&","\u0624");
       theMap.put("<","\u0625");
       theMap.put("}","\u0626");
       theMap.put("A","\u0627");
       theMap.put("b","\u0628");
```

```
        theMap.put("p","\u0629");
        theMap.put("t","\u062A");
        theMap.put("v","\u062B");
        theMap.put("j","\u062C");
        theMap.put("H","\u062D");
        theMap.put("x","\u062E");
        theMap.put("d","\u062F");
        theMap.put("*","\u0630");
        theMap.put("r","\u0631");
        theMap.put("z","\u0632");
        theMap.put("s","\u0633");
        theMap.put("$","\u0634");
        theMap.put("S","\u0635");
        theMap.put("D","\u0636");
        theMap.put("T","\u0637");
        theMap.put("Z","\u0638");
        theMap.put("E","\u0639");
        theMap.put("g","\u063A");
        theMap.put("_","\u0640");
        theMap.put("f","\u0641");
        theMap.put("q","\u0642");
        theMap.put("k","\u0643");
        theMap.put("l","\u0644");
        theMap.put("m","\u0645");
        theMap.put("n","\u0646");
        theMap.put("h","\u0647");
        theMap.put("w","\u0648");
        theMap.put("Y","\u0649");
        theMap.put("y","\u064A");

        theMap.put("P", "\u067E");
        theMap.put("J", "\u0686");
        theMap.put("V", "\u06A4");
        theMap.put("G", "\u06AF");

        // Numbers
        if (USING_LATIN_DIGITS) {
                theMap.put("0", "0");
                theMap.put("1", "1");
                theMap.put("2", "2");
                theMap.put("3", "3");
                theMap.put("4", "4");
                theMap.put("5", "5");
                theMap.put("6", "6");
                theMap.put("7", "7");
                theMap.put("8", "8");
                theMap.put("9", "9");
        } else {
                theMap.put("0", "\u0660");
                theMap.put("1", "\u0661");
                theMap.put("2", "\u0662");
                theMap.put("3", "\u0663");
                theMap.put("4", "\u0664");
                theMap.put("5", "\u0665");
                theMap.put("6", "\u0666");
                theMap.put("7", "\u0667");
                theMap.put("8", "\u0668");
                theMap.put("9", "\u0669");
        }
        // punctuation
        theMap.put(".", "."); //JAMES
```

```
        theMap.put(",", ","); // JAMES
        theMap.put("`","\u0670");
        theMap.put("{","\u0671");
        setUsingDiacritics(true);
    }

    /**
         * Adds/removes Buckwalter letters to/from the HashMap and their
         * respective Arabic script short vowel equivalents
         * @param using - (Boolean) True if using diacritics; false if not
         */
    public static void setUsingDiacritics(boolean using) {
                USING_DIACRITICS = using;
                if (USING_DIACRITICS) {
                        theMap.put("a","\u064E");
                        theMap.put("u","\u064F");
                        theMap.put("i","\u0650");
                theMap.put("~","\u0651");
                theMap.put("o","\u0652");
                theMap.put("F","\u064B");
                theMap.put("N","\u064C");
                theMap.put("K","\u064D");
                }
                else {
                        SHOW_UNTRANSLITERATED_CHARS = false;
                        SUBSTITUTE_CHAR = ""; // Suppress diacritics
                        theMap.remove("a");
                        theMap.remove("u");
                        theMap.remove("i");
                        theMap.remove("~");
                        theMap.remove("o");
                        theMap.remove("F");
                        theMap.remove("N");
                        theMap.remove("K");
                }
        }

    /**
         * Transliterates a Buckwalter string to Arabic string by
referencing the HashMap.
         * @param s - (String) The Buckwalter string to transliterate
         * @return result - (String) The transliterated Arabic script string
         */
   public static String transliterate(String s) {
    String ret = "";
    for (int i=0;i<s.length();i++) {
        String nextCh = theMap.get(Character.toString(s.charAt(i)));
        if (nextCh==null) {
                if (SHOW_UNTRANSLITERATED_CHARS) {
                        nextCh = Character.toString(s.charAt(i));
                } else {
                        nextCh = SUBSTITUTE_CHAR;
                }
        }
     ret +=nextCh;
    }
    return ret;
   }

   public static void main(String[] argv) {
```

```
        String inputBW = JOptionPane.showInputDialog("Enter a string in
Buckwalter transliteration");
        JOptionPane.showMessageDialog(null,
BwArMap.transliterate(inputBW));
 }
}
```

REFERENCES

Ababneh, A.H., Lu. J. and Xu, Q., 2016. Arabic Information Retrieval: A Relevancy Assessment Survey. In J. Gołuchowski, M. Pańkowska, C. Barry, M. Lang, H. Linger and C. Schneider (Eds.), *Information Systems Development: Complexity in Information Systems Development (ISD2016 Proceedings).* Katowice, Poland: University of Economics in Katowice.

Ababneh, M., Al-Shalabi, R., Kanaan, G. and AlNobani, A., 2012. Building an effective rule-based light stemmer for Arabic language to improve search effectiveness. *International Arab Journal of Information Technology,* 9(4), pp.368–372.

Abbasi, A., Chen, H. and Salem, A., 2008a. Sentiment analysis in multiple languages: feature selection for opinion classification in Web forums. *ACM Transactions on Information Systems,* 26(3), pp.1–34.

Abbasi, A., Chen, H., Thoms, S. and Fu, T., 2008b. Affect analysis of Web forums and blogs using correlation ensembles. *IEEE Transactions on Knowledge and Data Engineering,* 20(9), pp.1168–1180.

Abdelali, A., 2004. Localization in Modern Standard Arabic. *Journal of the American Society for Information Science and Technology,* 55(1), pp.23–28.

Abdelrahman, S.E., Mobarz, H., Farag, I. and Rashwan, M., 2014. Arabic phrase-level contextual polarity recognition to enhance sentiment Arabic lexical semantic database generation. *International Journal of Advanced Computer Science and Applications (IJACSA),* 5(10), pp.32–36.

Abdul-Mageed, M. and Diab, M., 2012a. WATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12),* Istanbul, Turkey: European Language Resources Association (ELRA).pp. 3907-3914.

Abdul-Mageed, M. and Diab, M., 2012b. Toward building a large-scale Arabic sentiment lexicon. In *Proceedings of the 6th International Global WordNet*, Matsue, Japan, pp.18–22.

Abdul-Mageed, M. and Diab, M., 2014. SANA: A large-scale multi-genre, multi-dialect lexicon for Arabic subjectivity and sentiment snalysis. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, pp.1,162–1,169.

Abdul-Mageed, M., Diab, M.T. and Korayem, M., 2011. 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*, Portland, Oregon: Association for Computational Linguistics, pp.587–591.

Abdul-Mageed, M., Diab, M. and Kubler, S., 2014. SAMAR: subjectivity and sentiment analysis for Arabic social media. *Computer Speech & Language*, 28(1), pp.20–37.

Abdul-Mageed, M., Kubler, S. and Diab, M., 2012. SAMAR: a system for subjectivity and sentiment analysis of Arabic social media. In *Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), ICC Jeju, Republic of Korea*, pp.1–19.

Adar, E., Zhang, L., Adamic, L. and Lukose, R., 2004. Implicit structure and dynamics of blogspace. *In Workshop on the Weblogging Ecosystem*, 13(1), pp. 16989–16995.

Ahmad, K., Gillam, L. and Cheng, D., 2005. Society grids. In *Proceedings of the UK e-Science all hands meeting 2005*, pp. 18-21.

AlGahtani, S., 2011. *Arabic named entity recognition: a corpus-cased study.* PhD, University of Manchester.

Alghamdi, H. and Selamat, A., 2015. Topic Modelling Used to Improve Arabic Web Pages Clustering. In *Proceedings of Cloud Computing (ICCC), International Conference on IEEE*, Riyadh, KSA: IEEE, pp. 1-6.

Alghamdi, H.M., Selamat, A. and Abdul Karim, N.S., 2014. Arabic web pages clustering and annotation using semantic class features. *Journal of King Saud University - Computer and Information Sciences, Special Issue on Arabic NLP*, 26(4), pp. 388–397.

Alhazmi, S., Black, W. and McNaught, J., 2013. Arabic SentiWordNet in relation to SentiWordNet 3.0. *International Journal of Computational Linguistics*, 4(1), pp. 1–11.

Alhazmi, S. and McNaught, J., 2013. Generating an Arabic sentiment corpus from social media. *Lancaster Workshop on Arabic Corpus Linguistics*, pp. 56–59.

Aliane, H., Guendouzi, A. and Mokrani, A., 2013. Annotating events, time and place expressions in Arabic texts. In *Proceedings of Recent Advances in Natural Language Processing*, Hissar, Bulgaria, pp. 25–31.

Almas, Y. and Ahmad, K., 2007. A note on extracting 'sentiments' in financial news in English, Arabic and Urdu. In *Proceedings of the 2nd Workshop on Computational Approaches to Arabic Script-based Languages (CAASL'07)*, Stanford: California, Linguistic Society of America, pp. 1–12.

Al-Sabbagh, R. and Girju, R., 2012. Yadac: Yet another dialectal Arabic corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12),* Istanbul, Turkey: European Language Resources Association (ELRA), pp. 2882–2889.

Al-Subaihin, A.A., Al-Khalifa, H.S. and Al-Salman, A.S., 2011. A proposed sentiment analysis tool for modern Arabic using human-based computing. In *Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services*, New York, NY, USA: ACM, pp. 543–546.

Altman, R., Bergman, C., Blake, J., Blaschke, C., Cohen, A., Gannon, F., Grivell, L., Hahn, U., Hersh, W., Hirschman, L., Jensen, L., Krallinger, M., Mons, B., O'Donoghue, S., Peitsch, M., Rebholz-Schuhmann, D., Shatkay, H. and Valencia, A., 2008. Text mining for biology – the way forward: opinions from leading scientists. *Genome Biology,* 9(2), pp. 1-15.

Ananiadou, S., Thompson, P., Kano, Y., McNaught, J., Attwood, T.K., Day, P.J., Keane, J., Jackson, D. and Pettifer, S., 2011. Towards interoperability of European language resources. *Ariadne (67),* Available at: http://www.ariadne.ac.uk/issue67/ananiadou-et-al/.

Andreevskaya, A. and Bergler, S., 2006. Mining WordNet for fuzzy sentiment: sentiment tag extraction from WordNet glosses. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 6, Trento, Italy, pp. 209–216.

Attia, M., 2008. *Handling Arabic morphological and syntactic ambiguity within the LFG framework with a view to machine translation.* PhD, University of Manchester.

Baccianella, S., Esuli, A. and Sebastiani, F., 2010. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10),* Irec, Malta, pp. 2200–2204.

Baeza-Yates, R., and Neto, B.R. (1999). *Modern information retrieval*. New York: ACM Press.

Balahur, A. and Balahur, P., 2009. What does the world think about you? Opinion mining and sentiment analysis in the social Web. *The Scientific Annals of "Alexandru Ioan Cuza" University of Iasi Communication Sciences*, pp. 101–110.

Balog, K., Mishne, G. and de Rijke, M., 2006. Why are they excited? Identifying and explaining spikes in blog mood levels. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 207–210.

Banry, R.K., 1997. *ALA-LC romanization tables: transliteration schemes for non-Roman scripts.* Washington: Cataloging Distribution Service, Library of Congress.

Baradaran, R., 2013. Developing Arabic event extraction system using rule-based method. In *Proceedings of the International Symposium on Advances in Science and Technology*, Qom, Iran: Department of Information Technology, University of Qom, pp. 1–10.

Beseiso, M., Ahmad, A. and Ismail, R., 2011. An Arabic language framework for semantic web. In *Proceedings of the International Conference on Semantic Technology and Information Retrieval*, Putrajaya, Malaysia: IEEE, pp. 7–11.

Biber, D., 1993. Representativeness in corpus design. *Literary and linguistic computing*, 8(4), pp. 243–257.

Bifet, A. and Frank, E., 2010. Sentiment knowledge discovery in Twitter streaming data. In *Proceedings of the 13th International Conference on Discovery Science*, Canberra: Springer-Verlag, pp. 1–15.

Bifet, A., Holmes, G., Pfahringer, B. and Gavaldà, R., 2011. Detecting sentiment change in Twitter streaming data. *Journal of Machine Learning Research – Proceedings Track 17*, pp. 5–11.

Binali, H., Potdar, V. and Wu, C., 2009. A state of the art opinion mining and its application domains. In *Proceedings of IEEE International Conference on Industrial Technology ICIT 2009*, Gippsland, VIC: IEEE, pp. 1–6.

Black, W., Elkateb, S. and Vossen, P., 2006. Introducing the Arabic WordNet project. In *Proceedings of the 3rd International WordNet Conference (GWC-06)*, Jeju Island, Korea, pp. 295–300.

Blei, D., Ng, A. and Jordan, M., 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3, pp. 993-1022.

Bollen, J., Mao, H. and Zeng, X.-J., 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), pp. 1–8.

Boulaknadel, S., 2008. Impact of Term-Indexing for Arabic Document Retrieval. In *Proceedings of International Conference on Application of Natural Language to Information Systems*, Springer Berlin Heidelberg, pp. 380-383.

Boulaknadel, S., Daille, B. and Aboutajdine, D., 2008a. A multi-word term extraction program for Arabic language. In *Proceedings of the 6th international Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, pp. 1485–1488.

Boulaknadel, S., Daille, B. and Aboutajdine, D., 2008b. Multi-word term indexing for Arabic document retrieval. In *Proceedings of IEEE symposium on Computers and Communications (ISCC'08)*, IEEE, pp. 869-873.

Bounhas, I. and Slimani, Y., 2009. A hybrid approach for Arabic multi-word term extraction. In *Proceedings of the International Conference on Language Processing and knowledge Engineering*, Dalian: IEEE, pp. 24–27.

Buckwalter, T., 2004. Issues in Arabic orthography and morphology analysis. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, Geneva: Association for Computational Linguistics, pp. 31–34.

Butler, C., 2004. Corpus studies and functional linguistic theories. *Functions of Language*, 11(2), pp. 147–186.

Carletta, J., 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2), pp. 249–254.

Choudhary, B. and Bhattacharyya, P., 2002. Text clustering using semantics. In *Proceedings of the 11th International World Wide Web Conference*, Honolulu, HI, USA: ACM, pp. 1–4.

Cohen, K. B., Johnson, H. L., Verspoor, K., Roeder, C. and Hunter, L., 2010. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics,* 11(1), pp. 1-10.

Conrad, J. G. and Schilder, F., 2007. Opinion mining in legal blogs. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, Stanford, CA: ACM, pp. 231–236.

Dagan, I. and Church, K., 1995. Termight: Identifying and translating technical terminology. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics (EACL) '95*, San Francisco, CA, USA: Morgan Kaufmann Publishers, pp. 34–40.

Daudé, J., Daudé, L. and Rigau, G., 2000. Mapping WordNets using structural information. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000),* Hong Kong: ACL, pp. 504–511.

Dave, K., Lawrence, L. and Pennock, D., 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web*, Budapest, Hungary: ACM, pp. 519–528.

Denecke, K., 2008. Using SentiWordNet for multilingual sentiment analysis. *ICDE Workshops IEEE Computer Society,* pp. 507–512.

Devitt, A. and Ahmad, K., 2007a. A lexicon for polarity: Affective content in financial news text. In *Proceedings of Language for Special Purposes (LSP'07),* Hamburg, pp. 1–3.

Devitt, A. and Ahmad, K., 2007b. Sentiment polarity identification in financial news: a cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague: Association for Computational Linguistics, pp. 984–991.

Diab, M., Al-Badrashiny, M., Aminian, M., Attia, M., Dasigi, P., Elfardy, H., Eskander, R., Habash, N., Hawwari, A. and Salloum, W., 2014. Tharwa: a large-scale dialectal Arabic - Standard Arabic - English lexicon. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, pp. 3782–3789.

Diab, M., Hacioglu, K. and Jurafsky, D., 2004. Automatic tagging of Arabic text: from raw text to base phrase chunks. In *Proceedings of HLT-NAACL 2004: Short Papers*, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 149–152.

Duwairi, R.M., Ahmed, N.A. and Al-Rifai, S.Y., 2015. Detecting sentiment embedded in Arabic social media – A lexicon-based approach. *Journal of Intelligent & Fuzzy Systems*, 29(1), pp. 107–117.

Duwairi, R.M. and Alshboul, M.A., 2015. Negation-aware framework for sentiment analysis in Arabic reviews. In *Proceedings of the 3rd International Conference on Future Internet of Things and Cloud*, Rome, Italy: IEEE Computer Society, pp. 731–735.

Eklundh, J. O. and Rosenfeld, A., 1978. Convergence Properties of Relaxation Labelling. Technical Report 701, Computer Science Center. University of Maryland.

Elarnaoty, M., AbdelRahman, S. and Fahmy, A., 2012. *A machine learning approach for opinion holder extraction in Arabic language.* arXiv preprint arXiv:1206.1011.

El-Beltagy, S.R. and Ali, A., 2013. Open issues in the sentiment analysis of arabic social media: A case study. In *Proceedings of the 9th International Conference on Innovations in Information Technology (IIT)*, IEEE, pp. 215–220.

El-Halees, A., 2011. Arabic opinion mining using combined classification approach. In *Proceedings of International Arab Conference on Information Technology (ACIT)*, Amman, Jordan: ACIT, pp. 1–8.

Elhawary, M. and Elfeky, M., 2010. Mining Arabic business reviews. In *IEEE International Conference on Data Mining Workshops*, Sydney, Australia: IEEE, pp. 1108–1113.

El-Kahky, A., Darwish, K., Saad Aldein, A., Abd El-Wahab, M., Hefny, A., and Ammar, W., 2011. Improved transliteration mining using graph reinforcement. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 1384–1393.

Elkateb, S., Black, W., Farwell, D., Vossen, P., Pease, A., and Fellbaum, C., 2006. *Arabic WordNet and the challenges of Arabic.* In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, London: The British Computer Society (BCS), pp. 15–25.

El-Khatib, K. and Badarenh, A., 2010. Automatic extraction of Arabic multi-word Term. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, Wisla: IEEE, pp. 411–418.

Elsawy, E., Mokhtar, M. and Magdy, W., 2014. TweetMogaz v2: identifying news stories in social media. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*,Shanghai: ACM (CIKM), pp. 1–3.

Elyaakoubi, M. and Lazrek, A., 2005. Arabic scientific e-document typography. In *Proceedings of the 5th International Conference on Human System Learning (ICHSL5)*, Marrakesh, Morocco, pp. 241–252.

Eskander, R. and Rambow, O., 2015. SLSA: a sentiment lexicon for Standard Arabic. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon: Association for Computational Linguistics, pp. 2545–2550.

Esuli, A. and Sebastiani, F., 2005. Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, Bremen: ACM, pp. 617–624.

Esuli, A. and Sebastiani, F., 2006. SENTIWORDNET: a publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on International Language Resources and Evaluation (LREC'06),* Genoa, Italy, pp. 417–422.

Färber, I., Günnemann, S., Kriegel, H., Kröger, P., Müller, E., Schubert, E., Seidl, T. and Zimek, A., 2010. On using class labels in evaluation of clusterings. In X.Z. Fern, I. Davidson, and J. Dy, eds. *MultiClust: discovering, summarizing, and using multiple clusterings.* ACM SIGKDD.

Farghaly, A. and Shaalan, K., 2009. Arabic natural language processing: challenges and solutions. *ACM Transactions on Asian Language Information Processing*, 8(4), pp. 1–22.

Farra, N., Challita, E., Abou Assi, R. and Hajj, H., 2010. Sentence-level and document-level sentiment mining for Arabic texts. In *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*, Sydney, NSW: IEEE, pp. 1114–1119.

Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstat, O., Rajman, M., Schler, Y. and Zamir, O.,1998. Text mining at the term level. *PKDD*, 1510, pp. 65–73.

Fleiss, J.L., Levin, B. and Paik, M.C., 2013. *Statistical methods for rates and proportions.* John Wiley & Sons.

Frantzi, K., Ananiadou, S. and Mima, H., 2000. Automatic recognition of multi-word terms. *International Journal of Digital Libraries*, 3(2), pp. 117–132.

Freitag, D., 2000. Machine learning for information extraction in informal domains. *Machine Learning*, 39(2), pp. 169–202.

Froud, H., Benslimane, R., Lachkar, A. and Ouatik, S., 2010. Stemming and similarity measures for Arabic documents clustering. In *Proceedings of the 5th International Symposium on I/V Communications and Mobile Network (ISVC),* Rabat: IEEE, pp.1–4.

Froud, H., Sahmoudi, I. and Lachkar, A., 2013. An efficient approach to improve Arabic documents clustering based on a new keyphrases extraction algorithm. *Computer Society*, pp. 243–256.

Gamon, M., 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th international conference on Computational Linguistics*, Geneva: Association for Computational Linguistics, pp. 1-7.

Gamon, M., Basu, S., Belenko, D., Fisher, D., Hurst, M., and König, A.C., 2008. Blews: using blogs to provide context for news articles. *In ICWSM*, pp. 60–67.

Graff, D., Maamouri, M., Bouziri, B., Krouna, S., Kulick, S., and Buckwalter, T., 2009. Standard Arabic morphological analyzer (SAMA) Version 3.1. *Linguistic Data Consortium, Catalog No.: LDC2010L01*.

Gross, M., 1997. The construction of local grammars. In Roche, E. and Schabès, Y. (eds), 1997. Finite State Language Processing. MIT Press, pp. 329–354.

Gruhl, D., Guha, R., Liben-Nowell, D. and Tomkins, A., 2004. Information diffusion through blogspace. In *Proceedings of the 13th International Conference on World Wide Web*, New York: ACM, pp. 491–501.

Hammad, M. and Al-Awadi, M., 2016. Sentiment Analysis for Arabic Reviews in Social Networks Using Machine Learning. In *Information technology: new generations.* Switzerland: Springer International Publishing, pp. 131–139.

Harris, Z., 1991. A Theory of Language and Information: A mathematical approach. Clarendon Press.

Hatzivassiloglou, V. and McKeown, K., 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of*

the Association of Computational Linguistics (ACL'97), Madrid: ACL, pp. 174–181.

Hetzron, R., 1997. *The Semitic languages.* London: Routledge.

Hirschman, L. and Mani, I., 2003. Evaluation. In R. Mitkov, ed. *Oxford handbook of computational linguistics.* Oxford: Oxford University Press, pp. 414–429.

Hu, M. and Liu, B., 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle: ACM, pp. 168–177.

Huang, A., 2008. Similarity measures for text document clustering. In *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008)*, Christchurch, New Zealand: NZCSRSC, pp. 49–56.

Ikeda, D., Fujiki, T. and Okumura, M., 2006. Automatically linking news articles to blog entries. In *American Association for Artificial Intelligence Spring Symposium: Computational Approaches to Analyzing Weblogs*, AAAI, pp. 78–82.

Jaccard, P., 1901. Etude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles,* 37, pp. 547–579.

Jain, A.K., Murty, M.N. and Flynn, P.J., 1999. Data clustering: a review. *ACM Computing Surveys (CSUR),* 31(3), pp. 264–323.

Java, A., Song, X., Finin, T. and Tseng, B., 2007. Why we Twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, San Jose, CA, USA: ACM, pp. 56–65.

Jensen, L. J., Saric, J. and Bork, P., 2006. Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics*, (9), pp. 119–129.

Justeson, J. and Katz, S., 1995. Technical terminol-ogy: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1), pp. 9–27.

Kano, Y., Baumgartner Jr., W., McCrohon, L., Ananiadou, S., Cohen, K., Hunter, L., and Tsujii, J., 2009. U-Compare: share and compare text mining tools with UIMA. *Bioinformatics*, 25, pp. 1997–1998.

Kano, Y., Miwa, M., Cohen, K.B., Hunter, L., Ananiadou, S., and Tsujii, J., 2011. U-Compare: a modular NLP workflow construction and evaluation system. *IBM Journal of Research and Development*, 55(3), pp. 11:1–11:10.

Kelaiaia1, A. and Merouani, H., 2016. Clustering with probabilistic topic models on Arabic texts: a comparative study of LDA and K-Means. *The International Arab Journal of Information Technology*, 13(2), pp. 332–338.

Kennedy, A. and Inkpen, D., 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22, pp. 110–125.

Kim, J.-D. and Tsujii, J., 2006. *Corpora and their annotation.* Boston and London: Artech House, pp. 179–211.

Kim, S.-M. and Hovy, E.H., 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 1-7.

Kiraz, G., 2002. Computational nonlinear morphology: with emphasis on Semitic languages. *Computational Linguistics*, 28(4), pp. 76–81.

Kok, S. and Brockett, C., 2010. Hitting the right paraphrases in good time. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.* Stroudsburg, PA, US: Association for Computational Linguistics, pp. 145–153.

Kolluru, B., Hawizy, L., Murray-Rust, P., Tsujii, J. and Ananiadou, S., 2011. Using workflows to explore and optimise named entity recognition for chemistry. *PLoS ONE*, 6(5), pp. e20,181.

König, A. C., Gamon, M. and Wu, Q., 2009. Click-through prediction for news queries. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: ACM, pp. 347–354.

Kontonatsios, G., K. I., Kolluru, B. and Ananiadou, S., 2011. Adding text mining workflows as web services to the BioCatalogue. In *Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences*, London, UK: ACM, pp. 50–57.

Kumar, R., Novak, J., Raghavan, P. and Tomkins, A., 2004. Structure and evolution of blogspace. *Com*, 47(12), pp. 35–39.

Kumar, S., Morstatter, F. and Liu, H., 2014. Analyzing Twitter data. In *Twitter data analytics.* New York: Springer, pp. 35–48.

Kwak, H., Lee, C., Park, H. and Moon, S., 2010. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web,* New York: ACM, pp. 591–600.

Larkey, S., Ballesteros, L. and Connell, E., 2007. Light stemming for Arabic information retrieval. In *Arabic omputational orphology*, New York: Springer, pp. 221–243.

Larsen, B. and Aone, C., 1999. Fast and effective text mining using linear-time document clustering. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, USA: ACM, pp. 16–22.

Leskovec, J., Adamic, L.A. and Huberman, B.A., 2007. The dynamics of viral marketing. *ACM Trans.Web,* 1(1), pp. 1–5.

Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N.S. and Hurst, M., 2007. Patterns of cascading behavior in large blog graphs. *SDM*, 7, pp. 551–556.

Linguistic Data Consortium, 2005a. *ACE (Automatic Content Extraction) Arabic annotation guidelines for events*, Philadelphia, PA: University of Pennsylvania.

Linguistic Data Consortium, 2005b. *ACE (Automatic Content Extraction) English Annotation guidelines for events,* Philadelphia, PA: University of Pennsylvania.

Lipiński, E., 2001. *Semitic languages: outline of a comparative grammar.* 2nd ed. vol. 80, Peeters Publishers.

Liu, B., 2010. Sentiment analysis and subjectivity. In N. Indurkhya and F. J. Damerau, ed. *Handbook of natural language processing,* 2, CRC Pres, pp. 627–666.

Liu, B., 2011. Opinion mining and sentiment analysis. In B. Liu, ed. *Web data mining.* Berlin, Heidelberg: Springer, pp. 459–526.

Liu, B., 2012. Sentiment analysis and *opinion* mining. *Synthesis Lectures on Human Language Technologies,* 5(1), pp.1-167.

Liu, B., 2015. *Sentiment analysis: mining opinions, sentiments and emotions.* Cambridge: Cambridge University Press.

Liu, B., Hu, M. and Cheng, J., 2005. Opinion observer: analyzing and comparing opinions on the Web. In *Proceedings of the 14th International Conference on World Wide Web*, Chiba, Japan: ACM, pp. 342–351.

Lo, Y. W. and Potdar, V., 2009. A review of opinion mining and sentiment classification framework in social networks. In *Proceedings of the Digital Ecosystems and Technologies, 3rd IEEE International Conference*, Istanbul, Turkey: IEEE, pp. 396–401.

Lu, Y., Mei, Q. and Zhai, C., 2011. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Information Retrieval,* 14(2), pp. 178–203.

Luo, Z., Osborne, M., Petrovic, S. and Wang, T., 2012a. Improving Twitter retrieval by exploiting structural information. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence,* Toronto, Ontario, Canada: AAAI, pp.648- 654.

Luo, Z., Osborne, M. and Wang, T., 2012b. *Opinion retrieval in Twitter.* In J.G. Breslin, N.B. Ellison, J.G. Shanahan and Z. Tufekci, eds. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media,* Dublin, Ireland: The AAAI Press.

Maamouri, M., Bies, A. and Kulick, S., 2006. Diacritization: a challenge to Arabic Treebank annotation and parsing. In *Proceedings of the Conference of the Machine Translation SIG of the British Computer Society,* Citeseer, pp. 35–47.

Manning, C.D., Raghavan, P. and Schütze, H., 2008. Evaluation in information retrieval. In *Introduction to information retrieval.* Cambridge: Cambridge University Press, pp. 151–175.

Manning, C.D., Raghavan, P. and Schütze, H., 2009. Text classification and Naive Bayes. In *Introduction to information retrieval.* Cambridge: Cambridge University Press, pp. 253–287.

Mansour, S., Sima'an, K. and Winter, Y., 2007. Smoothing a lexicon-based POS tagger for Arabic and Hebrew. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources,* Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 97–103.

Matsumoto, S., Takamura, H. and Okumura, M., 2005. Sentiment classification using word sub-sequences and dependency sub-trees. In *Proceedings of the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, Hanoi, Vietnam: Springer-Verlag, pp. 301–311.

Maybury, M., 2010. Collaborative analysis for information driven safeguards. *International Atomic Energy Agency Symposium on International Safeguards: Preparing for Future Verification Challenges, Vienna International Center*, 145, pp. 1–5.

McCallum, A.K., 2002. *MALLET: a machine learning for language toolkit*. Retrieved 23 April 2016 from http://mallet.cs.umass.edu.

McCallum, A., Corrada-Emmanuel, A. and Wang, X., 2005. Topic and role discovery in social networks. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, Edinburgh, Scotland, UK: IJCAI, pp. 786–791.

McGlohon, M., Akoglu, L. and Faloutsos, C., 2011. Statistical properties of social networks. In C.C. Aggarwal, ed. *Social network data analytics.* New York: Springer, pp. 17–42.

McLean, J., 2009. *State of the blogosphere.* Retrieved 8 August 2009 from http://technorati.com/blogging/article/.

Melville, P., Gryc, W. and Lawrence, R., 2009. *Sentiment analysis of blogs by combining lexical knowledge with text classification.* Paris, France: ACM, pp. 1275–1284.

Meryem, H., Ouatik, S.A. and Lachkar, A., 2014. A novel method for Arabic multi-word term extraction. *International Journal of Database Management Systems (IJDMS)*, 6(3), pp. 53–67.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K., 1990. Introduction to WordNet: an on-line lexical database*. *International Journal of Lexicography*, 3(4), pp. 235–244.

Mishne, G. and de Rijke, M., 2006. A study of blog search. In *Proceedings of the 28th European Conference on IR Research (ECIR 2006),* London, UK: Springer Berlin Heidelberg, 3936, pp. 289–301.

Mitchell, C., 2012. The Australian. Retrieved 1 August 2012 from http://www.theaustralian.com.au/news/opinion.

Mittal, A. and Goel, A., 2011. *Stock prediction using Twitter sentiment analysis.* Stanford, CA: Stanford University.

Mobarz, H., Rashwan, M. and AbdelRahman, I., 2011. Generating lexical resources for opinion mining in Arabic language automatically. In *Proceedings of the 11th Conference on Language Engineering (ESOLE),* Cairo, Egypt.

Montejo-Raez, A., Martinez-Camara, E., Martin-Valdivia, M. and Ureña Lopez, L.A., 2014. Ranked WordNet graph for sentiment polarity classification in Twitter. *Computer Speech & Language*, 28(1), pp. 93–107.

Moreno-Ortiz, A. and Hernández, C., 2013. Lexicon-based sentiment analysis of Twitter messages in Spanish. *Procesamiento del Lenguaje Natural*, 50, pp. 93–100.

Mourad, A. and Darwish, K., 2013. Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA'13,* Atlanta, Georgia: Association for Computational Linguistics, pp. 55–64.

Nakagawa, H., and Mori, T. (2002). A simple but powerful automatic term. In *Proceedings of the 2nd International Workshop on Computational Terminology (COMPUTERM2002)*, Taiwan: Association for Computational Linguistics, pp. 29–35.

Niwattanakul, S., Singthongchai, J., Naenudorn, E. and Wanapu, S., 2013. Using of Jaccard Coefficient for keywords similarity. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Hong Kong: IMECS, pp. 380–384.

Ohana, B. and Tierney, B., 2009. Sentiment classification of reviews using SentiWordNet. In *Proceedings of the 9th IT&T Conference*, Dublin: Dublin Institute of Technology, pp. 13–xx.

Okazaki, N. and Ananiadou, S., 2006. Building an abbreviation dictionary using a term recognition approach. *Bioinformatics*, 22, pp. 3089–3095.

Orimaye, S.O., 2011. Sentence-level contextual opinion retrieval. In *Proceedings of the 20th International Conference Companion on World Wide Web*, Hyderabad, India: ACM, pp. 403–408.

Pak, A. and Paroubek, P., 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Irec, Malta: LREC, pp. 1320–1326.

Pang, B. and Lee, L., 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, Barcelona, Spain: Association for Computational Linguistics, pp. 271–278.

Pang, L. and Lee, L., 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2), pp. 1–135.

Pang, B., Lee, L. and Vaithyanathan, S., 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 79–86.

Phelan, O., McCarthy, K. and Smyth, B., 2009. Using Twitter to recommend real-time topical news. In *Proceedings of the third ACM Conference on Recommender Systems*, New York: ACM, pp. 385–388.

Piao, S., Ananiadou, S., Tsuruoka, Y., Sasaki, Y., and McNaught, J., 2007. Mining opinion polarity relations of citations. *International Workshop on Computational Semantics (IWCS)*, pp. 366–371.

Pourrajabi, M., Moulavi, D., Campello, R.J., Zimek, A., Sander, J., and Goebel, R., 2014. Model selection for semi-supervised clustering. In *Proceedings of the 17th International Conference on Extending Database Technology (EDBT)*, Athens, Greece: EDBT, pp. 331–342.

Prabowo, R. and Thelwall, M., 2009. Sentiment analysis: a combined approach. *Journal of Informetrics*, 3(2), pp. 143–157.

Rao, D. and Ravichandran, D., 2009. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, Athens, Greece: Association for Computational Linguistics, pp. 675–682.

Refaee, E. and Rieser, V., 2014. An Arabic Twitter corpus for subjectivity and sentiment analysis. In *Proceedings of the 9th International Conference*

on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland: LREC, pp. 2268–2273.

Řehůřek, R. and Sojka, P., 2011. *Gensim-Python framework for vector space modelling.* Brno, Czech Republic: Masaryk University.

Richards, J., Landgrebe, D. and Swain, P., 1981. On the accuracy of pixel relaxation labelling. *IEEE Transactions on Systems,* Man and Cybernetics, 11(4), pp.303-309.

Rosen-zvi, M., Griffiths, T., Steyvers, M. and Smyth, P., 2004. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, Arlington, Virginia, US: AUAI Press, pp. 487–494.

Rushdi-Saleh, M., Martín-Valdivia, M.T., Ureña-López, L.A. and Perea-Ortega, J.M., 2011. OCA: Opinion corpus for Arabic. *Journal of the American Society for Information Science and Technology,* 62(10), pp. 2045–2054.

Salameh, M., Mohammad, S. M. and Kiritchenko, S., 2015. Sentiment after translation: a case-study on arabic social media posts. In *Proceedings of the North American Chapter of Association of Computational Linguistics*, Denver, Colorado, USA: Association for Computational Linguistics, pp. 767–777.

Saleh, L. and Al-Khalifa, H., 2009. AraTation: an Arabic semantic annotation tool. In *Proceedings of the 11th International Conference on Information Integration and Web-Based Applications and Services,* Lumpur, Malaysia: ACM, pp. 447–451.

Salton, G., 1989. *Automatic text processing.*New York: Addison-Wesley.

Salvetti, F., Lewis, S. and Reichenbach, C., 2004. Automatic opinion polarity classification of movie reviews. *Colorado Research in Linguistics,* 17(1), pp.1-15.

Sayyadi, H., Hurst, M. and Maykov, A., 2009. Event detection and tracking in social streams. In *Proceedings of the 3rd International ICWSM Conference*, San Jose, California, US: AAAI, pp. 311–314. San Jose, California

Schluter, P.M. and Harris, S.A., 2006. Analysis of multilocus fingerprinting data sets containing missing data. *Molecular Ecology Note,* 6, pp. 569–572.

Schouten, K. and Frasincar, F., 2016. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering,* 28(3), pp. 813–830.

Seeber, F., 2008. Citations in supplementary information are invisible. *Nature,* 451(7181), pp. 887– 888.

Shaalan, K., Allam, A. and Gomah, A., 2003. Towards automatic spell checking for Arabic. In *Proceedings of the 4th Conference on Language Language Engineering (ELSE'03)*, Cairo: Egyptian Society of Language Engineering, pp. 240–247.

Shaban, K., 2009. A semantic approach for document clustering. *Journal of Software*, 4(5), pp. 391–404.

Sharma, J. and Vyas, A., 2010. *Twitter sentiment analysis.* Indian Institute of Technology unpublished report.

Shimodaira, H., 2015. Text classication using Naive Bayes. INFR08009 *Informatics 2B: Algorithms, data dtructures, learning*. University of Edinburgh, UK.

Shoukry, A. and Rafea, A., 2012. Sentence-level Arabic sentiment analysis. In *Collaboration Technologies and Systems (CTS), 2012 International Conference* Denver, Colorado, USA: IEEE, pp. 546–550.

Sinclair, J., 2005. Corpus and text: basic principles. In M. Wynne, ed. *Developing Linguistic Corpora: a Guide to Good Practice.* Oxford: Oxbow Books, pp. 1–16.

Smeeton, N., 1985. Early history of the Kappa statistic. *Biometrics*, 41, pp. 795–795.

Stavrianou, A. and Chauchat, J.-H., 2008. Opinion mining issues and agreement identification in forum texts. *Atelier FODOP 08*, pp. 51–58.

Steinbach, M., Karypis, G. and Kumar, V., 2000. A comparison of document clustering techniques. In *KDD Workshop on Text Mining*, 400(1), pp. 525–526.

Steyvers, M. and Griffiths, T., 2007. Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis and W. Kintsch, eds. *Handbook of latent semantic analysis*, Erlbaum: Psychology Press, pp.424-440.

Stone, A., 2001. *Transliteration of Indic scripts: how to use ISO 15919*, Geneva: International Organization for Standardization (ISO).

Strapparava, C. and Mihalcea, R., 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, Fortaleza, Brazil: ACM, pp. 1556–1560.

Strehl, A., Ghosh, J. and Mooney, R., 2000. Impact of similarity measures on web-page clustering. In *Proceedings of AAAI Workshop on AI for Web Search*, Austin, TX, USA: AAAI, pp. 58–64.

Szabó, G. and Huberman, B.A., 2010. Predicting the popularity of online content. *Communications of the ACM*, *53*(8), pp.80–88.

Taboada, M., 2016. Sentiment analysis: an overview from linguistics. *Annual Review of Linguistics*, 2, pp. 325–347.

Taboada, M., Tofiloski, M., Stede, M., Brooke, J., and Vollz, K., 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 1(1), pp. 1–42.

Takama, Y., Matsumura, A. and Kajinami, T., 2006. Visualization of news distribution in blog space. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Washington, DC: IEEE Computer Society, pp. 413–416.

Takamura, H., Inui, T. and Okumura, M., 2004. Extracting emotional Polarity of words using spin model. In *Proceedings of the Joint Workshop of Vietnamese Society of AI, SIGKBS-JSAI, ICS-IPSJ and IEICE-SIGAI on Active Mining*, Hanoi, Vietnam, pp. 207–212.

Thelwall, M., 2006. Bloggers during the London attacks: top information sources and topics. In *Proceedings of the World Wide Web 2006 Workshop on the Weblogging*, 8, pp. 1–8.

Thelwall, M., 2013. *Heart and soul: sentiment strength detection in the Social Web with SentiStrength.* Wolverhampton: University of Wolverhampton, Statistical Cybermetrics Research Group.

Thelwall, M., Buckley, K. and Paltoglou, G., 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1), pp. 163–173.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A., 2010. Sentiment in short strength detection informal text. *Journal of the Association for Information Science and Technology (JASIST)*, 61(12), pp. 2544–2558.

Thompson, P., Kano, Y., McNaught, J., Pettifer, S., Attwood, T. K., Keane, J., and Ananiadou, S., 2011. Promoting interoperability of resources in META-SHARE. In *Proceedings of the Workshop on Language Resources, Technology and Services in the Sharing Paradigm*, Chiang Mai, Thailand: IJCNLP, pp.50–58.

Tounsi, L. and Genabith, J. V., 2010. Arabic parsing using grammar transforms. In *Proceedings of LREC 2010 – the 7th Conference on International Language Resources and Evaluation*, Valletta, Malta: LREC, pp. 1986–1989.

Tsagkias, M., de Rijke, M. and Weerkamp, W., 2009. Predicting the volume of comments on online news stories. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, Hong Kong: ACM. pp. 1765–1768.

Tsagkias, M., de Rijke, M.D. and Weerkamp, W., 2011. Linking online news and social media. In *Proceedings of the 4th ACM International*

Conference on Web Search and Data Mining, Hong Kong: ACM, pp. 565–574.

Tsytsarau, M. and Palpanas, T., 2012. Survey on mining subjective data on the web. *Data Mining Knowledge Discovery*, 24(3), pp. 478–514.

Turney, P.D., 2002. 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*, Philadelphia, PA: Association for Computational Linguistics, pp. 417–424.

Turney, P. and Littman, M., 2003. Measuring praise and criticism: inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4), pp. 315–346.

UN Development Programme, U., 2009. *Arab human development report 2009: challenges to human security in the Arab countries.* New York, USA: UNDP/RBAS.

Vechtomova, O., 2010. Facet-based opinion retrieval from blogs. *Information processing & management*, 46(1), pp. 71–88.

Verma, A., Kaur, I. and Arora, N., 2016. Comparative analysis of information extraction techniques for data mining. *Indian Journal of Science & Technology* , 9(11), pp. 1–18.

Versteegh, K., 2001. *The Arabic language.* Edinburgh: Edinburgh University Press.

Viera, A.J. and Garrett, J.M., 2005. Understanding interobserver agreement: the kappa statistic. *Family Medicine,* 37, pp. 360–363.

Wagner, D., Venezky, R. and Street, B., 1999. *Literacy: An International Handbook.* Boulder, CO: Westview Press.

Weber, G., 1997. The world's 10 most influential languages. *Language Today,* 3, pp. 12–18.

Wilson, T., Wiebe, T. and Hoffmann, P., 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver: Association for Computational Linguistics, pp. 347–354.

Wilson, T., Wiebe, J. and Hwa, R., 2004. Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings of the 19th National Conference on Artifical Intelligence (AAAI), 4*, San Jose, California, US: Association for the Advancement of Artificial Intelligence, pp. 761–769.

Witten, I.H., Frank, E. and Hall, M.A., 2011. *Data mining: practical machine learning tools and techniques.* 3rd ed. San Francisco: Morgan Kaufmann.

Yang, K., Yu, N. and Zhang, H., 2007. WIDIT in TREC-2007 Blog Track: combining lexicon-based methods to detect opinionated blogs. In *Proceedings of the 16th Text Retrieval Conference (TREC 2007),* pp. 1–12.

Zaidan, O.F. and Callison-Burch, C., 2014. Arabic dialect identification. *Computational Linguistics,* 40(1), pp. 171–202.

Zhang, W., Y. C. and Meng, W., 2007. Opinion retrieval from blogs. In *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management*, Lisbon: ACM, pp. 831–840.

Zhao, Y. and Karypis, G., 2002. Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the 11th International Conference on Information and Knowledge Management*, McLean, VA, USA: ACM, pp. 515–524.

Zhao, Y. and Karypis, G., 2004. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55(3), pp. 311–331.

Zhu, T.T., Zhang, F.X. and M., L., 2013. ECNUCS: a surface information-based system description of sentiment analysis in Twitter in the SemEval-2013 (Task 2). In *Proceedings of SEM 2013: The Second Joint Conference on Lexical and Computational Semantics, 2*, Atlanta, Georgia, USA: Association for Computational Linguistics, pp. 408–413.

Zweigenbaum, P., Demner-Fushman, D., Yu, H. and Cohen, K.B., 2007. Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics*, 8(5), pp. 358–375.