Journal of Intelligent & Fuzzy Systems 29 (2015) 107–117
DOI:10.3233/IFS-151574
IOS Press

Detecting sentiment embedded in Arabic social media – A lexicon-based approach

R.M. Duwairi[a,*], Nizar A. Ahmed[b] and Saleh Y. Al-Rifai[b]
[a]*Department of Computer Information Systems, Jordan University of Science and Technology, Irbid, Jordan*
[b]*Department of Computer Science, Jordan University of Science and Technology, Irbid, Jordan*

Abstract. Sentiment analysis aims at extracting sentiment embedded mainly in text reviews. The prevalence of semantic web technologies has encouraged users of the web to become authors as well as readers. People write on a wide range of topics. These writings embed valuable information for organizations and industries. This paper introduces a novel framework for sentiment detection in Arabic tweets. The heart of this framework is a sentiment lexicon. This lexicon was built by translating the SentiStrength English sentiment lexicon into Arabic and afterwards the lexicon was expanded using Arabic thesauri. To assess the viability of the suggested framework, the authors have collected and manually annotated a set of 4400 Arabic tweets. These tweets were classified according to their sentiment into positive or negative tweets using the proposed framework. The results reveal that lexicons are helpful for sentiment detection. The overall results are encouraging and open venues for future research.

Keywords: Sentiment analysis, unsupervised learning, text mining, Arabic text, opinion mining

1. Introduction

Data is available and is easily accessible for anyone connected to the internet. Huge amounts of data are uploaded on the internet on a daily basis. Todays' challenges for companies, stakeholders, organizations, or individuals are not to access data but to extract knowledge from data or make sense of data [7, 28, 29]. One form of knowledge that is embedded in data is peoples' opinions about products and services. Discovering such opinions or sentiments is vital for companies and organizations. Often companies' success depends on how well their products or services are perceived by customers and therefore companies invest heavily in building tools that extract opinions or sentiments about their products or services.

Opinion mining or sentiment analysis is the field of science that is interested in extracting opinions embedded in customers' reviews [1, 42, 49]. Customers' opinions could be expressed in a structured way such as using stars to rate a product as is commonly used with Amazon [10] and the Movie Database [45]. By the same token, people use free text to express their opinions or sentiments. The latter is common in social media channels such as Facebook [24] and Twitter [65]. Sentiment analysis may deal with extracting the polarity of the text (positive, negative or neutral) [1, 2, 23], determining whether text contains bad or good news [37], finding whether a candidate is likely to win or to loose [35], finding the stated outcome of a drug: improvement or death [47], finding whether a person supports or opposes a post [12], rank entities such as cars or hotels based on their associated opinionated data [27], and many more. Sentiment Analysis has been extensively studied in the literature for the English language. By comparison, relatively few works have targeted sentiment analysis in Arabic text [e.g. 1, 2, 5, 8, 20–23, 56].

There are several granularities for sentiment analysis. A popular work is to determine whether a text is subjective or objective [50]. Another common work targets the valence or polarity of the text (i.e. positive or negative)

*Corresponding author. R.M. Duwairi, Department of Computer, Information Systems, Jordan University of Science and Technology, Irbid 22110, Jordan. Tel.: +962 2 720 1000; Fax: +962 2 720 1077; E-mail: rehab@just.edu.jo.

[19]. A third category deals with finding the strength of an emotional state in text [63]. A much deeper and therefore challenging analysis is to detect the exact emotion covered by a text such as "happy", "sad" and "angry" [13, 70]. Lastly, the most challenging analysis is to find users' intensions or arguments [69].

Generally speaking, there are two approaches for detecting sentiment in text [16, 33]. The first one relies on linguistic resources such as dictionaries and lexicons [e.g. 17, 31, 32, 40, 43, 48, 60, 62, 71]. The second one is based on machine learning [e.g. 1, 4–6, 15, 21, 22]. Some researchers have combined the previous two approaches [39]. The major challenges for building lexicons is that these are very hard to build manually and they are domain dependent but they do not require an annotated dataset to detect sentiment. Machine learning (Classification in particular), on the other hand, employs a labeled dataset for sentiment detection. The major challenge here is to build the annotated dataset for the classification task. Transfer learning can help in training classifiers on new domains using labelled data from another domain [73].

Sentiment analysis is hard to detect for many reasons; one reason is that people use different writing styles to express their opinions. A second reason is that sentiment is context dependent [68]. Often, polarity bearing words may become neutral when considering the context. For example, in Arabic, nouns with positive polarities are used as person names such as the word (حكيم *Hakeem*); which means *wise* in English. حكيم as an adjective indicates positive sentiment but as a person name it is neutral (i.e. it has no sentiment). Therefore, to accurately detect sentiment, the context of the word or phrase should be explored. Also, people opinions change over time. A much bigger challenge in sentiment analysis comes from the fact that people usually express their opinions in a comparative manner, and people tend to express their positive and negative reviews in the same sentence. Consider for example the following sentence: *the movie*'*s idea was great but the actors*' *performance was modest*. Irony is a way of communication where the speaker says something and means the opposite with the absence of negation markers [52, 59]. Detecting sentiment in ironic sentences is a real challenge to computational algorithms as well as to humans because sentiment is implicit in irony and does not have clear markers.

Research on Arabic sentiment analysis is isolated and scattered. There is no benchmark datasets that researchers can compare their results against. There are no standard sentiment lexicons that researchers could benefit from. It is common among researchers working with Arabic language to build their own datasets or lexicons for the purpose of their study and the story ends there. These local datasets or lexicons are seldom made public.

Arabic is a morphologically rich language and this creates challenges for researchers working on Natural Language Processing (NLP), text mining or machine learning. Even though text mining and machine learning do not require deep language analysis as it is the case with NLP, still simple tasks such as tokenization and stemming are nontrivial tasks in Arabic. For example, in English a sentence always starts with a capital letter and ends with a period. Arabic, by comparison, does not use capital letters and it does not have strict punctuation regime and thus a sentence could end with a comma, semicolon, period, colon, and so on. Further, it is common in Arabic to co-join sentences via (و) *wa* and (ف) *fa* [25]. Normalization is another challenge for researchers; some letters have the same shape except for an added dot (.), Hamza (ء) or Mada (~). For example is it common to find the following variations of letter *alif*: (ا أ آ إ) which corresponds to letter "A or a" in English. Arabic often neither incorporates vowels nor adheres to the proper inclusion of marks above or below letters. To overcome this, Arabic letters are normalized. For example, it is common to normalize the several forms of *alif* mentioned above to *plain alif* (ا).

Stemming in Arabic is not equivalent to the removal of suffixes or prefixes. During stemming a word is reduced to its three-letter root; this means in addition to dealing with suffixes and prefixes, we have to deal with infixes as well. In addition to that, some words in Arabic have 4-letter or 5-letter roots. A less aggressive approach, called light-stemming [9], has been used by Arabic researchers instead of stemming. In light stemming, common prefixes and suffixes are removed. For example, the root of (وكتابهم) '*and their book*' is (كتب) '*write*' and light-stem of the same word is (كتاب) '*book*'.

Intensifiers have great impact on sentiment analysis as well. Arabic has rich rules for intensification that vary from using specific words such as *very* (جداً) to the repetition of complete phrases. For example, *very happy* may be expressed as (سعيد جداً), or (جداً سعيد). Note that the intensifier, in Arabic, could come before or after the word. There are special patterns or forms of words that serve the purpose of intensification or mitigation. In sentiment analysis, researchers have to take

care of valence shifters such as negation as these usually flip the sentiment of a word. Again Arabic has a rich set of negation letters and rich rules for using them. A nice study about negation recognition in medical fields for the English language can be found in [53]. Arabic language has dialects; it is common for users in social media channels to write using dialectical or colloquial Arabic. Colloquial Arabic extends the vocabulary and grammatical rules of Modern Standard Arabic (MSA). In social media channels, it is also common for Arab bloggers to write Arabic text using Latin letters (Known as Arabizi [11, 18]). As it can be seen, dealing with Arabic text is a nontrivial task. The current research has provided solutions for some of the above issues.

This research introduces a novel framework for sentiment analysis in Arabic reviews which relies on a sentiment lexicon. Thus this work falls under unsupervised learning. The core idea of this research is to build a semantic lexicon that will aid in determining the polarity of tweets written in Arabic. The seed of the lexicon was borrowed from SentiStrength [57] and it was expanded by using thesauri. The lexicon is embedded in a framework that determines the polarity of tweets written in Arabic. Every tweet is tokenized into terms, then every term is assigned a weight equals to 1 if the term is indicated as positive in the lexicon; assigned a weight equals to -1 if the term is indicated as negative in the lexicon; or assigned a weight equals zero if the it is inexistent in the lexicon. Afterwards, a tweet is said to carry positive sentiment if the summation of its terms' weights is greater than zero. By comparison, the tweet is said to carry negative sentiment if the summation of its terms' weights is less than zero. Lastly, a tweet is considered neutral if the summation of its terms' weights is equal to 0. Two experiments were carried out on the dataset of tweets: the first one does not employ stemming and the second one stems the tweets as part of the processing. The results obtained reveal that stemming does improve sentiment analysis in the Arabic language. The best precision was equal to 0.70 and the best recall was equal to 0.46. These results are close to the best results obtained when employing unsupervised learning for sentiment detection in tweets. See Section 4 for more details.

The rest of this paper is organized as follows: Section 1 has introduced the current work. Section 2, by comparison, provides some background information on the topic and lists some related work. Section 3, on the other hand, introduces our framework. Section 4 describes the experimentation setup and analyzes the results of lexicon-based sentiment analysis. Finally, Section 5 draws the conclusions of this work and highlights future work.

2. Background and related work

2.1. Background

Lexicon-based sentiment detection is a class of algorithms that attempts to determine the polarity of a text or review by combining the polarity of words or phrases which appear in that text. These rely on a lexicon or dictionary which includes words or phrases with their sentiment orientation. It is believed that words have prior polarity regardless of context. For example, the word *hate* expresses negative sentiment while the word *love* indicates positive sentiment when considered without reference to the contexts in which they are used. Of course, this assumption is not entirely true when contexts are taken into consideration. A simple example would be the use of negation which reverses the prior polarity of words. A second example would be when sentiment bearing words are used as person names and thus become neutral when the context is considered. Intensification and mitigation affect the degree of sentiment. For instance, "*very good*" and "*good*" are examples of positive sentiment bearing words but with different intensities.

Unsupervised or lexicon-based methods for sentiment analysis consist of two major tasks, namely: building the lexicon and creating the parser that will utilize the lexicon. For the first task, several researchers have resorted to the manual annotation of words; others have used manual annotation to create a *seed* for the lexicon and afterwards the seed was extended by using linguistic rules, WordNet [26, 41], posting queries to search engines or using more elegant algorithms such as the work reported in [51]. For the second task, the simplest approach is based on tokenizing the text into words and afterwards these words are looked up in the lexicon to determine their prior polarity. The overall polarity of the text is usually determined by summing the weights obtained from the lexicon. A more elegant approach uses linguistic information, such as POS, when determining the prior polarity of tokens or words such as handling negation.

On the other hand, several researchers have approached sentiment analysis as a classification task and thus they have borrowed classifiers from the machine learning field and applied them to sentiment

analysis and opinion mining. Supervised leaning when used for sentiment analysis gives hire accuracies when compared with lexicon-based approaches but it comes at the cost of manual labeling of datasets.

2.2. Related work

The following paragraphs present selected works that either have used supervised or unsupervised sentiment analysis. The works also have been applied to several languages: mainly English and a few on Arabic.

SAMAR [2] is a two stage classifier which first distinguishes subjective sentences from objective ones written in Arabic. Secondly, it determines the polarity of subjective sentences. SVM light was used for both stages. The authors studied the effect of adding morphology knowledge to the sentences and its effects on the classifier accuracy. The dataset that they have experimented with consists of 8940 sentences.

The work reported in [19] relies on a simple set of rules applied at the syntactic level. It argues that the syntactic stand of a word is related to its sentiment. Therefore, the parse tree of the sentence is employed to derive rules that determine the polarity of the sentence.

Steinberger et al. [60] build a sentiment dictionary in multiple languages. The procedure that they have used consists of collecting sentiment words in English and Spanish. After these lists are approved by humans, they are translated into a third language, say French, via machine translation. Only common words in the two translated lists are kept as sentiment words in French. The idea here is to overcome errors of translation by translating from two source languages to one destination language (called triangulation). These sentiment lists or dictionaries can be used to enhance multilingual sentiment analysis.

Taboada et al. [61] developed a word-based method for detecting sentiments in English text. They named it SO-CAL (semantic calculator). They manually built a lexicon of words and phrases with their sentiments from -5 to 5 (-5 means extremely negative and 5 means extremely positive). The current version of SO-CAL includes, in addition to adjectives, nouns, verbs and adverbs. SO-CAL handles negation, intensification and mitigation in intelligent ways. Their experimentation proved that SO-SAL performs well on several domains and that SO-CAL is in harmony with human judgment as well.

Tufis and Stefanescu [64] developed a framework for the annotation of WordNet 3.0 [26] based on differential semantics. They used word senses for all words of WordNet not only adjectives. Using word senses enabled them to assign different annotations for different senses of the same word. Their work is valuable for researchers working on sentiment analysis based on lexicons of valence. This work is comparable to SentiWordNet [8], and WordNet Affect [66]. The work reported in [38] presents a lexicon model for annotating words (verbs, nouns and adjectives) to be used in sentiment analysis applications.

The authors in [30] propose a framework for sentiment analysis which first relies on a sentiment lexicon and un-annotated data to train a classifier. Then the initial classifier is applied to un-annotated text. Documents with high classification confidence are used to extract sentiment features which are subsequently used to train a second classifier. This second classifier is then applied to the test data.

The authors in [72], on the other hand, introduce a strategy to predict the semantic orientation of words without the online support of the internet. This strategy first builds a semantic orientation model (semantic vector space). Afterword's, a classifier is trained to identify the semantic orientation of words and phrases. The results of their empirical evaluation outperform other known methods.

Mourad and Darwish [44] proposed a method for classifying tweets written in Arabic. In total, there were 2300 tweets involved in their experiments. Their approach is a supervised one that uses Naïve Bayes for sentiment detection. However, a sentiment lexicon was used to enhance the set of features that the classifier would use during training. In particular, the MPQA [67] English lexicon was translated using Bing online Machine Translation tool [14] into Arabic words in addition to the ArabSenti lexicon [3]. The resultants set of features was extended using a graph reinforcement algorithm which runs over phrase tables generated using Moses [36] to extract synonyms of features. In their experimental setup, several variations of features were used such as stem or root prior sentiment, stem part-of-speech tags, stem bi-grams (to capture negation), counts of POS tags, and whether the stem is strong-subjective or weak-subjective. The accuracy reported for the baseline ArabSenti lexicon was 76.6% (for subjectivity detection); and it did not improve when the expanded lexicon was used. For the polarity detection, the baseline accuracy was 80.5%, and the expanded-lexicon accuracy was actually reduced to 80.0%. This limited study may indicate that the use of words prior sentiment supersedes the case were this is extended with more words such as synonyms.

Shoukry and Refae in [58] have worked on a tweet dataset that consists of 1000 tweets (500 are positives and 500 are negative). They worked on sentence-level sentiment analysis since tweets length is restricted to 140 characters. Though their work lacks handling the neutral class and employs a small corpus, they explored the direction of Arabic dialects and appended some Egyptian words alongside the Modern Standard Arabic. For preprocessing phase, they applied unigrams and bigrams and concluded that there was no difference in the results. The approach followed, in their paper, was corpus-based (supervised approach), where SVM and NB were used for polarity classification. The results show that SVM outperformed NB in sentiment analysis with accuracy reached 72.6% regardless to the feature extraction technique used (either unigrams or bigrams).

As it can be seen from the above related work, too many researchers have targeted sentiment analysis. The approaches vary from being lexicon-based to machine learning based. Some researchers have combined machine learning with lexicons to improve the results. The advantages the lexicons have over supervised learning are that lexicons do not require annotated dataset to build their models and that they perform well on several domains.

3. The proposed framework

This section explains the two major components of the current work, namely, the lexicon and the unsupervised sentiment detection module. Subsection 3.1 explains the procedure that was used to create the lexicon while Subsection 3.2 explains how this lexicon was used in detecting sentiments embedded in Arabic tweets.

3.1. Arabic sentiment lexicon

This lexicon is a major component of the current framework. The following steps explain how this lexicon was built:

a) The English lexicon, SentiStrength [57], was used as a seed to build the Arabic sentiment lexicon. The version of SentiStrength that was used in this article consists of 300 words. These words were translated to Arabic words using English-Arabic Dictionary.
b) The polarity of every word was determined by two individuals. This polarity is expressed as -1 to indicate a negative sentiment or as 1 to indicate a positive sentiment. Every word is assigned only one sentiment weight.
c) The lexicon was extended by including a list of synonyms for every word in the seed list. Sakhr dictionary [55] was used to generate the synonym lists. The synonyms of a word have a polarity equals to the polarity of that word. i.e. synonyms of a positive word are considered positive and synonyms of a negative word are considered negative. The length of every synonym list is either two or three words.
d) Khoja [34] stemmer was used to stem the words which are included in the lexicon so far (The result of step c above). The extracted roots are also appended to the lexicon. A root of a positive word is positive and vice versa. Due to the fact that Arabic is morphologically rich and it uses diacritics, several words with opposite sentiments may have the same three letter root. For example, the word, "تلاعب" "tlAEb"[1] which means *manipulate* has the three-letter root "لعب" "lEb" is a negative word in Arabic and thus it has a polarity label equals (-1). By comparison, the word "يلعب" "ylEb" which means *play* has the three-letter root "لعب" "lEb" is a positive word in Arabic and thus has the polarity label (1). As another example, the word "تمييز" "tmyyz" which means *discrimination* has the three-letter root "ميز" 'myz" and it is a negative word. On the other hand, the word "إمتياز" "<mtyAz" which means *excellent* has the three-letter root"ميز" and it is a positive word. The previous two examples show that words with conflicting sentiment orientation may end up having the same root. Such words create ambiguity for any algorithm that attempts to detect polarity of words. It is an interesting research topic to investigate the relationship between Arabic morphology and sentiment analysis. The current research opts to remove such ambiguous words manually from the lexicon.
e) In Arabic, letter alif "ا" may be combined with letter Hamza "ء" to produce two forms of the letter alif, namely: "أ إ". It is custom to normalize such letters to the plain alif "ا". In this work we did not follow this approach; we simply repeated the words with several representations of the alif Hamza.

[1] The Bukwalters' transliteration system is used here to represent Arabic alphabet using Roman characters.

Table 1
A sample of the emoticons list

Symbol	Label	Symbol	Label
%-(	−1	8\	−1
%-)	1	8c	−1
(-:	1	:#	−1
(:	1	:'(	−1
(∧ ∧)	1	:'-(	−1
(∧-∧)	1	:(	−1
(∧.∧)	1	:)	1

f) Emoticons (http://en.wikipedia.org/wiki/Emoticon) with their sentiment polarity were appended to the lexicon. Emoticons are international signs which are often used by social media users to express their feelings. For example, :) indicates positive sentiment and :(expresses negative sentiment when read from left to right. For Arabic language, which is read and written from right to left, emoticons also need special consideration. Table 1 shows a sample of the list of emoticons that were used in this work.

After applying step (a) to step (f) above, the lexicon ended up having 2376 entries; 1776 negative ones and 600 positive ones. A word which does not belong to the lexicon is considered neutral, for the current work, and thus has the label 0. Figure 1 shows the pseudo code of the process that is used to create the Arabic sentiment lexicon. Table 2, shows a sample of the lexicon.

3.2. Unsupervised sentiment detection

In unsupervised sentiment detection there is no need for a training corpus – the polarity of a tweet is simply determined by using the tokens of the tweet and the sentiment lexicon. In the current work, the polarity of a given tweet is calculated in the following manner (Refer to Fig. 2):

Table 2
A sample of the Arabic sentiment lexicon

Word	Label	Word	Label
ابادة	−1	ابتهاج	1
ابتذال	−1	ابدع	1
ابتذل	−1	ابطال	−1
ابتز	−1	ابعاد	−1
ابتزاز	−1	ابعد	−1
ابتسم	1	ابكم	−1
ابتلاء	−1	ابله	−1
اجهاد	−1	اهتز	−1

- Lexicon is empty
- Initialize lexicon (*seed* = 300 words from SentiStrength)
- Translate seed to Arabic using English to Arabic Dictionary
- Assign polarity labels to elements of *seed*
- Append *seed* with labels to Lexicon
- Expand *seed* with lists of synonyms using Sakhr Thesaurus
- Assign polarity labels to synonyms' lists
- Append labeled synonym lists to Lexicon
- Stem all entries in Lexicon
- Label all stems
- Append stems to Lexicon
- Append emoticons with their labels to Lexicon

Fig. 1. Pseudo code for creating the lexicon.

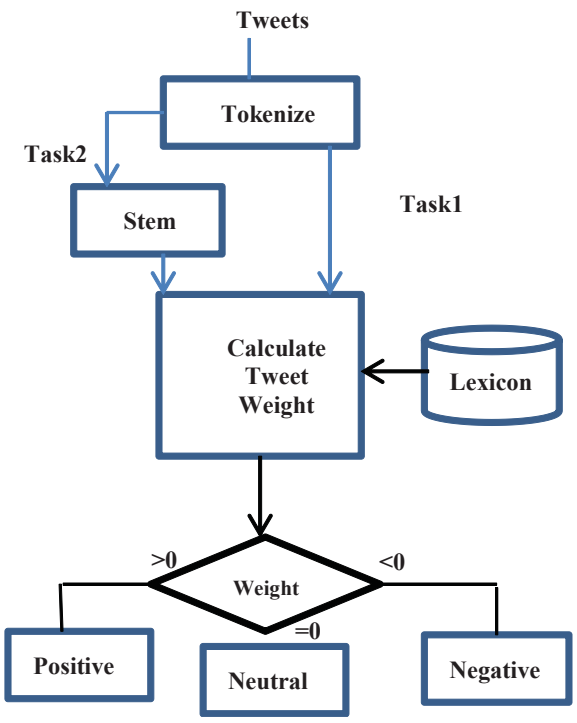

Fig. 2. Architecture of the framework of lexicon-based sentiment analysis.

1. Load a tweet from the database.
2. Extract unigrams of the tweet. Keep two versions of every tweet, namely, original words (Task 1 in Fig. 2) and stemmed words (Task 2 in Fig. 2).
3. Consult the Arabic Sentiment Lexicon to assign a label for every unigram in the tweet (generated from step 2 above). (−1) for negative unigrams, (1) for positive unigrams and (0) for neutral unigrams (the ones that do not exist in the lexicon).

4. In the case where a given unigram is a negation word, then reverse the polarity of the next unigram in the sequence.
5. Calculate the overall polarity of the tweet by summing the scores of all the unigrams in the tweet.
6. Assign the label "positive" to the tweet if its summation (weight), which is calculated in step 5 above, is greater than zero. Assign the label "negative" to the tweet if its overall weight is less than zero. Finally assign the label "neutral" if tweet weight equals to zero. Figure 2 depicts the overall architecture of lexicon-based sentiment analysis.

As an example of using our suggested framework, consider the tweet: "احب علوم الحاسوب" which is translated to "*I love computer science*" in English. After tokenizing the previous tweet and without stemming, we end up with the following tokens: "احب", "علوم", and "الحاسوب". The prior polarities of the previous tokens, as indicated by the lexicon, are: "احب":1, "علوم":0 and "الحاسوب":0. The summation of 1, 0, and 0 is 1 and therefore the previous tweet expresses positive sentiment. Consider as a second example the tweet "لا احب علوم الحاسوب" which is translated to "*I do not like computer science*" in English. After tokenizing the second tweet and without applying stemming, we end up with the following tokens: "علوم", "احب", "لا", and "الحاسوب". Note that "لا" is a negation word that was used to reverse the meaning of "احبlove". Thus the tokens with their prior sentiment polarities after handling negation become like: "لا احب":−1, "علوم":0 and "الحاسوب":0. The summation of the polarity labels of the previous tokens is −1 and hence the second tweet is considered a negative one.

4. Experimentation and result analysis

4.1. Dataset

In order to test the performance of the suggested framework, 4400 tweets, which are written in Arabic by users in response to certain events, were collected. These tweets were manually annotated with their sentiment: positive or negative. Every tweet was annotated by two independent annotators. The final label, that a tweet gets, is the label that both annotators agree on. In case of ties; i.e. one annotator says the current tweet is positive while the second annotator says the current tweet is negative; a third annotator was called to break out the tie. 3213 tweets were labeled as positive and 1187 tweets were labelled as negative. It is important to notice that the current framework is an unsupervised one, which means that it does not really need a labelled dataset and as such the current dataset was used to test the proposed framework. The human labels represent the absolute truth about the sentiments of the tweets. If the predicted label of a tweet matches the human assigned label, then this is considered a successful hit for the framework and vice versa.

4.2. Assessment framework

It is common for classification tasks to use precision, recall, error rate or accuracy to judge the quality of the produced results. These can be defined using the confusion matrix listed in Table 3:

Considering the confusion matrix presented in Table 3, the above metrics are calculated as shown in the next four formulae:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$\text{Error-rate} = \frac{FP + FN}{TP + FP + TN + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Table 3
Confusion matrix for calculating accuracy

Human		Computer	
		YES	NO
	YES	**TP** (true positives): number of tweets that both the human and computer agree to belong to the current class	**FN** (false negatives): number of tweets that the human says they belong to the current class but the computer says they do not belong to that class
	NO	**FP** (false positives): number of tweets that the computer program classifies them to belong to the current class while the human says they do not belong to that class.	**TN** (true negatives): number of tweets that both the human and computer agree that they do not belong to the current class

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

4.3. Lexicon-based experimentation

In this set of experiments, the sentiment lexicon was used to determine the polarity of tweets. In Experiment 1 all the 4400 tweets were classified using the unsupervised sentiment detection framework. During preprocessing, all stopwords, except negation words, were removed. The lexicon was used to determine the polarity of the remaining words. In this experiment, the words of the tweets were not stemmed. Thus, the research question of Experiment 1 is to judge the quality of the proposed framework when stemming is not used. A second experiment on the same set of tweets (Called Experiment 2) was carried out. Here, Khoja [34] stemmer was used to stem the tokens of the tweets before they were fed to the sentiment detection framework. The idea of Experiment 2 is to determine the effect of stemming on sentiment analysis.

Table 4 shows the values of the four assessment metrics that we have used. As Table 4 clearly shows, stemming enhances the performance (accuracy, precision and recall). This is because the likelihood of finding a stem in the dictionary is higher than finding the original word. For example, the verb "love" in English has many forms in Arabic such as "يحبون يحبان تحبان أحب نحب يحب تحب" etc. All have the same root "حبب". In the case of stemming, any of the several previous forms would be stripped to its three-letter root and if that root exists in the lexicon then sentiment can be calculated. Contrary to that, in the case where the original word is used, if the exact form of the word is not found in the lexicon then the polarity of that word cannot be determined. The precision for Experiment 1 was equal to 0.45 and for Experiment 2 it was equal to 0.70. Recall, on the other hand, was equal to 0.24 for Experiment 1 and was equal to 0.46 in Experiment 2. Accuracy was equal to 0.23 in Experiment 1 and was equal to 0.46 for Experiment 2. Finally, Error-Rate was equal to 0.77 in Experiment 1 and 0.54 in Experiment 2. These numbers shows that stemming does improve the precision, recall and overall accuracy; and it reduces error rate. The Results of Experiment 2 are close to the best results obtained when utilizing unsupervised learning for sentiment detection from reviews. For example, SemEval 2013 Task 2 was dedicated to sentiment analysis in Twitter [46]. Annotated datasets were provided to the participants. The task consists of two subtasks; the first one, deals with phrase-level polarity identification and the second subtask deals with tweet/message level polarity identification. Most of the competing systems where classification systems (i.e. they used machine learning for sentiment analysis) and only a few were semi-supervised systems. The best F1 measure for subtask one was 88.93% and the F1 score of the semi-supervised system was equal to 85.5%. F1 score is the harmonic mean of precision and recall. The highest F1 score for subtask 2 was equal to 69.02% for supervised systems and the best semi-supervised system F1 score was equal to 62.55%. Even though direct comparison between these systems and ours is not possible because different datasets were used, we claim that our unsupervised system gave very competing results. Specifically, the F1-score of our unsupervised system is equal to 55.51% for Experiment 2 (2*(Precision*Recall)/(Precision+Recall)). This value is not far from 62.02% given by the best performing *semi-supervised* system for subtask 2 which is the closest to our task. Also, SemEval 2014 Task 9 targeted sentiment analysis [54]. The best F1 score was equal to 70.96 which is not that high considering these systems are supervised ones. The authors of this paper anticipate that increasing the size of the lexicon will improve the results. They also anticipate that adding dialects lexicons will also improve the results. It is worth noting that tweets are short, 140 characters at most, which means a tweet may not include enough words to determine its polarity.

Table 4
Accuracy of the lexicon-based approach

Assessment	Exp. 1	Exp. 2
Precision	0.45	0.70
Recall	0.24	0.46
Accuracy	0.23	0.46
Error-Rate	0.77	0.54

5. Conclusions and future work

This paper has introduced a novel framework for sentiment analysis in Arabic tweets. The core of this framework is a sentiment lexicon. This lexicon was built by translating the terms of SentiStrength from English to Arabic and afterwards it was expanded using Arabic thesauri. This lexicon consists of 2376 entries: 1777 negative entries and 600 positive entries. The lexicon was subsequently employed to determine the polarity

of tweets. The overall sentiment of a tweet equals the summation of its respective terms' weights. The sentiment of every term was determined with the help of the lexicon. Positive terms were assigned a weight equals to 1; while negative terms were assigned a weight equals to −1. Neutral terms were assigned a weight equals to 0. A tweet is considered positive if its terms' summation is greater than zero. On the other hand, a tweet is considered negative if its terms' summation is less than zero. A tweet is considered neutral if the summation of its terms' weight is equal to 0.

Also, a dataset of 4400 tweets was collected and annotated to test the viability of the proposed framework. The human labels were considered the true labels. The lexicon was operated in two modes. The first mode, the terms were used as is without preprocessing. In the second mode, the terms were stemmed using Khoja [34] stemmer.

The results reveal that stemming does improve the overall accuracy. The obtained results were also close and comparable to other works that target sentiment analysis written in Arabic. This work is far from over. In the future, it can be extended in many ways such as increasing the size of the lexicon, employing a dialect lexicon, and extending the text of the tweet by using synonym lists.

References

[1] A. Abbasi, C. Hsinchun and S. Arab, Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums, *ACM Transactions on Information Systems (TOIS)* **26**(3) (2008), 1–34.

[2] M. Abdul-Mageed, S. Kubler and M. Diab, SAMAR: A system for subjectivity and sentiment analysis of Arabic Social Media. 3rd Workshop on Computational Approaches for Subjectivity and Sentiment Analysis WASSA, Satellite Workshop, Jeju, Koera, 2012.

[3] M. Abdul-Mageed, M. Diab and M. Korayem, Subjectivity and sentiment analysis of modern standard Arabic, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short papers-Vol 2*, 2011, pp. 587–591.

[4] K. Ahmad, D. Cheng and Y. Almas, Multi-lingual Sentiment Analysis of Financial News Streams. *Proceedings of the 1st International Workshop on Grid Technology for Financial Modeling and Simulation*, Palermo, Italy, 2006.

[5] K. Ahmad and Y. Almas, Visualizing Sentiments in Financial Texts. *Proceedings of the Ninth International Conference on Information Visualization*, Washington, USA, 2005, PP. 363–368.

[6] G. Alec, R. Bhayani and L. Huang, Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 2009, pp. 1–12.

[7] S. Al-Harbi, A. Almuhareb, A. Al-Thubaity, M.S. Khorsheed and A. Al-Rajeh, Automatic Arabic text classification. *Proceedings of the 9th International Conference on the Statistical Analysis of Textual Data*, Lyon-, France, 2008.

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

[9] M. Aljlayl and O. Frieder, On Arabic search: Improving the retrieval effectiveness via a light stemming approach. *Proceedings of the ACM 11th Conference on Information and Knowledge Management*, NewYork: ACM Press, 2002, pp. 340–347.

[10] Amazon, http://www.amazon.com, lasted accessed 11 September 2014.

[11] Arabizi, http://thehashemitekingdom.blogspot.com/2011/03/partial-arabizi-dictionary-or-dialect.html, last accessed 11 September 2014.

[12] M. Bansal, C. Cardie and L. Lee, The power of negative thinking: Exploiting label disagreement in the min cut classification framework. *Proceedings of the International Conference on Computational Linguistics (COLING)*, Poster paper, 2008, pp. 15–18.

[13] H. Binali, C. Wu and V. Potdar, Computational Approaches for Emotion Detection in Text. *Proceedings of 4th IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2010)*, Dubai, United Arab Emirates, 2010, pp. 172–177.

[14] Bing Online Machine Translator, http://www.bing.com/translator, Last accessed on 11 September 2014.

[15] E. Boiy and M.F. Moens, A machine leaning approach to sentiment analysis in multilingual web texts, *Information Retrieval* **12**(5) (2009), 526–558.

[16] Q. Cao, M.A. Thompson and Y. Yu, Sentiment analysis in decision sciences research: An illustration to IT governance, *Decision Support Systems* **54** (2013), 1010–1015.

[17] K. Denecke, Are SentiWordNet Scores Suited for Multi-domain Sentiment Classification? *In Fourth International Conference on Digital Information Management (ICDIM)*, Ann Arbor, MI, 2009, pp. 33–38.

[18] M. Diouri, Arabizi: A Contemporary Style of Arabic slang, http://www.mitpressjournals.org/doi/pdf/10.1162/desi.2008.24.2.39, last accessed 25 March 2014.

[19] A. Duric and F. Song, Feature selection for sentiment analysis based on content and syntax model, *Decision Support Systems* **53** (2012), 704–711.

[20] R. Duwairi, R. Marji, N. Shaban and S. Rushaidat, Sentiment Analysis in Arabic Tweets. *Proceedings of the 5th International Conference on Information and Communication Systems*, Irbid, Jordan, 2014.

[21] A. El-Halees, Arabic Opinion Mining Using Combined Classification Approach. *Proceedings of the International Arab Conference on Information Technology (ACIT)*, Riyadh, Saudi Arabia, 2011.

[22] M. Elhawary and M. Elfeky, Mining Arabic Business Reviews. *Proceedings of the IEEE International Conference on Data Mining*, Mountain View, USA, 2010, pp. 1108–1113.

[23] M. El-Orfali, *Experimenting with Arabic Opinion Mining*, M.Sc. Thesis, Qatar University 2012.

[24] Facebook, http://www.facebook.com, last accessed 11 September 2014.

[25] A. Farghaly and K. Shaalan, Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Languages Information Processing* **8**(4) (2009), Article 14.

[26] C. Fellbaum, *WordNet: An Electronic Lexical Database*, Academic Press, Cambridge, MA, 1998.

[27] K. Ganesan and C. Zhai, Opinion-based entity ranking, *Information Retrieval* **15**(2) (2012), 116–150.
[28] R. Gopal, J.R. Marsden and J. Vanthienen, Information mining-reflections on recent advancements and the road ahead in data, text, and media mining, *Decision Support Systems* **51** (2011), 727–731.
[29] F. Harrag, E. El-Qawasmeh and P. Pichappan, Improving Arabic text categorization using decision trees. Networked Digital Technologies. NDT'09. *First International Conference on*, 2009, pp. 110–115.
[30] Y. He and D. Zhou, Self-training from labeled features for sentiment analysis, *Information Processing and Management* **47** (2011), 606–616.
[31] J. Kamps, M.J. Marx, J. Robert and M. De Rijke, Using WordNet to measure semantic orientations of adjectives. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Vol. IV, 2004, pp. 1115–1118.
[32] J. Jiao and Y. Zhou, Sentiment polarity analysis based multi-dictionary. *Physics Procedia* **22** (2011), 590– 596.
[33] R. Jonathon and J. Carroll, Weakly supervised techniques for domain-independent sentiment classification. *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, 2009, pp. 45–52. ACM.
[34] S. Khoja and R. Garside, Stemming Arabic Text. Computing Department, Lancaster University, Lancaster, UK, 1999 http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps
[35] S.M. Kim and E. Hovy, Automatic detection of opinion bearing words and sentences, *Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, Jeju Island, Korea, 2005.
[36] P. Koehn, et al., Moses: Open source toolkit for statistical machine translation, *Annual Meeting for the Association of Computational Linguistics*, 2007.
[37] L.W. Ku, L.Y. LY, T.H. Wu and H.H. Chen, Major topic detection and its application to opinion summarization. *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR)*, Salvador, Brazil 2005.
[38] I. Maks and P. Vossen, A lexicon model for deep sentiment analysis and opinion mining applications, *Decision Support Systems* **53** (2012), 680–688.
[39] P. Melville, G. Wojciech and R.D. Lawrence, Sentiment analysis of blogs by combining lexical knowledge with text classification. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009, pp. 1275–1284. ACM.
[40] T. Mike, K. Buckley and G. Paltoglou, Sentiment strength detection for the social web, *Journal of the American Society for Information Science and Technology* **63**(1) (2012), 163–173.
[41] G.A. Miller, WordNet: A lexical database for english, *Communications of the ACM* **38**(11) (1995), 39–41.
[42] A. Montoyo, P. Martinez-Barco and A. Balahur, Subjectivity and sentiment analysis: An overview of the current area and envisaged developments, *Decision Support Systems* **53** (2012), 675–679.
[43] A. Moreo, M. Romero, J.L. Castro and J.M. Zurita, Lexicon-based comment-oriented news sentiment analyzer system, *Expert Systems with Applications* **39** (2012), 9166–9180.
[44] A. Mourad and K. Darwish, Subjectivity and sentiment analysis of modern standard Arabic microblogs. *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Atlanta, Georgia, USA, 2013, pp. 55–64.
[45] The Movie Database, http://www.imdb.com, last accessed 11 September 2014.
[46] P. Nakov, Z. Kozareva, A. Ritter, S. Rosenthal, V. Stoyanov and T. Wilson. SemEval 2013: Task 2: Sentiment anlysis in Twitter. *Second Joint Conference on Lexical and Computational Semantics, Volume 2: Seventh Innternational Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, 2013, pp. 312–320.
[47] Y. Niu, X. Zhu, J. Li and G. Hirst, Analysis of polarity information in medical text. *Proceedings of the American Medical Informatics Association*, 2005, Annual Symposium.
[48] B. Ohana and B. Tierney, Sentiment Classification of Reviews using SentiWordNet. *In 9th IT & Conference*, Dublin, Ireland, 2009.
[49] B. Pang and L. Lee, Opinion mining and sentiment analysis, *Foundations and Trends Information Retrieval* **2**(1–2) (2008), 1–135.
[50] B. Pang and L. Lee, A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 2004.
[51] G. Qiu, B. Liu, J. Bu and C. Chen, Opinion word expansion and target extraction through double propagation, *Computational Linguistics* **37**(1) (2010), 9–27.
[52] A. Reyes and P. Rosso, On the difficulty of automatically detecting Irony: Beyond a simple case of negation, *In: Knowledge and Information Systems* **40**(3) (2014), 595–614.
[53] L. Rokach, R. Romano and O. Maimon, Negation recognition in medical narrative reports, *Information Retrieval* **11**(6) (2008), 499–538.
[54] S. Rosenthal, P. Nakov, A. Ritter and V. Stoyanov, SemEval-2014 Task 9: Sentiment Analysis in Twitter. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, 2014, pp. 73–80.
[55] Sakhr, http://www.sakhr.com/. Last accessed 11 September 2014.
[56] M.R. Saleh, M.T. Martin-Valdivia, L.A. Urena-Lopez and J.M. Perea-Orteg, A OCA: Opinion corpus for Arabic, *Journal of the American Society for Information Science and Technology* **62**(10) (2011), 2045–2054.
[57] SentiStrength, http://sentistrength.wlv.ac.uk/SentStrength_Data/, Last accessed 11 September 2014.
[58] A. Shoukry and A. Rafea, Sentence-level Arabic sentiment analysis. *Proceedings of Collaboration Technologies and Systems (CTS)*, 2012, pp. 546–550.
[59] A. Sigar and Z. Taha, A contrastive study of Ironic expressions in english and Arabic, *College of Basic Education Researchers Journal* **12**(2) (2012), 795–817.
[60] J. Steinberger, et al., Creating sentiment dictionaries via triangulations, *Decision support Systems* **53** (2012), 689–694.
[61] M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede, Lexicon-based methods for sentiment analysis, *Computational Linguistics* **37**(2) (2011), 267–307.
[62] S. Tan and Q. Wu, A random walk algorithm for automatic construction of domain-oriented sentiment lexicon, *Expert Systems with Applications* **38** (2011), 12094–12100.
[63] M. Thelwall, K. Buckly, G. Paltoglou, D. Cai and A. Kappas, Sentiment strength detection in short informal text, *Journal of the American Society for Information Science and Technology (JASIST)* **61**(12) (2010), 2544–2558.

[64] D. Tufis and D. Stefanescu, Experiments with a differential semantics annotation for WordNet 3.0, *Decision Support Systems* **53** (2012), 695–703.
[65] Twitter, http://www.twitter.com, last accessed 11 September 2014.
[66] A. Valitutti, C. Strapparava and O. Stock, Developing affective lexical resources, *Psychology Journal* **2**(1) (2004), 61–83.
[67] J. Wiebe, T. Wilson and C. Cardie, Annotating expressions of opinions and emotions in language, *Language Resources and Evaluation* **39** (2005), 165–210.
[68] T. Willson, J. Wiebe and P. Hoffmann, Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis, *Computational Linguistics* **35**(3) (2009), 9p. 399–433, MIT Press Cambridge, MA, USA.
[69] T. Wilson, Fine-grained subjectivity and sentiment analysis: Recognizing the intensity, polarity and attitudes of private states, Ph. D. Thesis, 2008.
[70] C. Wu, Z. Chuang and Y. Lin, Emotion recognition from text using semantic labels and separable mixture models, *ACM Transactions on Asian Language Information Processing (TAIP)* **5**(2) (2006), 165–183.
[71] H. Xia, J. Tang, H. Gao and H. Liu, Unsupervised Sentiment Analysis with Emotional Signals." Rio de Janeiro, Brazil ACM 978-1-4503-2035-1/13/05, 2013.
[72] T. Xu, Q. Peng and Z. Cheng, Identifying the semantic orientation of terms using S-HAL for sentient analysis, *Knowledge-based Systems* **35** (2012), 279–289.
[73] D. Zhang, L. Si and V.J. Rego, Sentiment detection with auxiliary data, *Information Retrieval* **15**(3-4) (2012), 373–390.