Discovering semantic and sentiment correlations using huge corpus of short informal Arabic language text

Editor(s):
Solicited review(s):
Open review(s):

Al-Osaimi, Salha[a], Badruddin Khan, Muhammad[b]
salha@ccis.imamu.edu.sa, badruddin@ccis.imamu.edu.sa
Dept. of Information Systems
Imam Muhammad ibn Saud Islamic university Riyadh, KSA

Abstract. Semantic and Sentiment analysis have received a great deal of attention over the last few years due to the important role they play in many different fields, including marketing, education, and politics. Social media has given tremendous opportunities for researchers to collect huge amount of data as input for their semantic and sentiment analysis. Using twitter API, we collected around 4.5 million Arabic tweets and used them to propose a novel automatic unsupervised approach to capture patterns of words and sentences of similar contextual semantics and sentiment in informal Arabic language at word and sentence levels. We used Language Modeling (LM) model which is statistical model that can estimate the distribution of natural language in effective way. The results of experiments of proposed model showed better performance than classic bigram and latent sematic analysis (LSA) model in most of cases at word level. In order to handle the big data, we used different text processing techniques followed by removal of the unique words based on their relevance to problem.

Keywords: Informal Arabic, Big Data, Sentiment analysis, Opinion Mining (OM), semantic analysis, bigram model, LSA model, Twitter

1. Introduction

The last decade has seen a huge increase in the number of internet users in Middle East. This growth has helped in enriching the amount of Arabic content on website. There are wide numbers of users that use the social networks. They use social media in order to share various kinds of resources, express their opinions, thoughts, and messages in real time [1]. Since most of users use informal Arabic in the world of social media, the task of semantic and sentiment analysis becomes more sophisticated. Different Arabic Dialects are another challenge [2]. One of the main challenges is the limited number of researches that focus on the informal Arabic sentiments analysis. This motivated us to focus on the problems that exist in the realm of informal Arabic semantic and sentiment analysis thus encouraging the researchers to participate more in this field. Sentiment analysis, also called opinion mining, is the field of study that extracts and analyzes people's opinions about products, services, individuals, event, issues, to name but a few categories [3][4][5]. An opinion can be a positive or negative or neutral sentiment, attitude, emotion, or appraisal. For small corpus of data, it is possible to use humans for annotation but for big data, the formulation of training and test data is very expensive and almost impossible. Although a tweet is small piece of data but to annotate them when they are millions followed by application of machine learning techniques and then analyzing classification models to understand the polarity of different words is pretty difficult and expensive job.

This work is neither related to supervised learning nor it use existing semantic resources like Arabic WordNet due to informal nature of Arabic text in tweets. Our proposed approach does not depend on the syntactic structure of tweets, it extracts patterns from the contextual semantic and sentiment similarities between words in a given tweet corpus. Contextual semantics are based on the proposition that meaning can be extracted from words co-occurrences [6]. The LM model gives a probability distribution—or P(s)—over sequences of words (w_i). The goal of LM is to build a statistical model that can estimate the distribution of natural language in effective way [7] [8]. It uses a number of types of matrices, such as the unigram, bigram, and trigram. The bigram matrix is sometimes referred to as the word co-occurrence matrix [9][10]. In this study, we use a bigram matrix method for document representation. In the bigram matrix, each row represents a word (w_i), and each column represents the first preceding word (w_j) of w_i where j = i-1. Each cell gives the co-occurrence frequency (a_{ij}) of the word sequence w_j w_i in the corpus [9] [11][12].

The paper is organized in few sections to describe further details of our work to extract semantic and sentiments from the huge corpus of Arabic tweets. Section 2 outlines the related work done in this area. In section 3, describes the methodology of our work. In Section 4, discuss the experiments results. Finally, in the brief Section 5 we will make concluding remarks.

2. Related Work

This section provides a literature review for the field of sentiment and semantic analysis, focusing mainly on informal Arabic language.

2.1 Sentiment analysis in informal Arabic language

Duwairi, Marji, Sha'ban, and Rushaidat look at Arabic dialects, Arabism and emoticons. At the normalization stage, they add new step, which is to convert Arabic dialect to Modern Standard Arabic (MSA) by mapping dialect words on to MSA stems. Their study applied three different classifiers: Support Vector Machines (SVM), Naive Bayes (NB), and KNN. The accuracy of the SVM was found to be slightly lower than that of NB [13]. Both [14] and [2] have produced applications for Arabic sentiment analysis in order to classify Arabic tweets. They used the SVM and Naive Bayes classifiers, and also tried to classifiers together. Itani, Hamandi, Zantout, and Elkabani have studied the use of informal Arabic on Facebook. Their corpus contained eight different dialects; namely, Lebanese, Egyptian, Syrian, Iraqi, Libyan, Algerian, Tunisian, Sudanese, and Saudi. They built a classifier model using the Naive Bayes classifier. Accuracy was measured by comparing human and automatic classification results [15]. Other researches have focused upon the lexicon-based approach, which, typically, is used less often in Arabic sentiment analysis because of the low number of existing Arabic sentiment lexicons. The main challenge here is in building lexicons for informal words, as [1] [16] [17] and [18]. These studies encourage researchers to contribute more extensively to the field. El-Beltagy and Ali (2013) use the semantic orientation approach (SO) to determine Arabic Egyptian polarities, using two data sets: a Twitter data set and a Comments data set. The experiment showed that SO is effective, especially within the context of Twitter [16]. One of the latest sentiment analysis studies has been conducted by [17]. They analyzed three constructed lexicons, one manual and two automatic, designing a lexicon-based model for sentiment analysis. The result of performance is 74.6% is very encouraging. However, some interesting research has been undertaken that uses semantic analysis methods with the aim of improving the sentiment model. Unfortunately, in terms of our context, these studies focus on the English language. Saif, He, and Alani demonstrate the importance of using semantic features when determining the positive or negative sentiments in tweets. In their study, they used both tweet- and entity-level sentiment analysis [19]. They also propose a further study capturing the patterns of word with similar contextual semantics and sentiments in tweets [6]. [20] used a vector space model that learns word representations in order to capture semantic and sentiment words.

2.2 Semantic analysis in informal Arabic language

This section offers an overview of some studies that have applied semantic analysis to Arabic language data sets. The amount of Arabic language documents available online is increaseing with time. It is difficult for researchers to handle huge volumes of relevant texts documents. For this reason, Arabic document clustering is an important task for achieving perfect outcomes with Information Retrieval (IR) programs, thus satisfying a researcher's needs. Froud, Lachkar, and Ouatik have proposed a method for improving document categorization by using the topic map method, based on a method similar to document clustering. Their method was found to be quite effective for clustering documents, when compared with evaluation methods involving human beings [21]. Other study has sought to group the semantic features of Arabic web pages, clustering them based on their similarities, with the help of the Arabic VerbNet lexicon. The researchers collected a corpus from the archives of digital Arabic newspapers [22]. Other researchers propose the use of an Arabic language model for speech recognition and machine translation tasks [23] [24]. Notably, Sarikaya et al. introduced the joint morphological-lexical language model (JMLLM), which takes advantage of Arabic morphology, being designed for inflected languages in general (and Arabic in particular). They have used their system to conduct experiments into dialectal Arabic (Iraqi Arabic specifically). The results showed that JMLLM offers encouraging improvements when compared with base-line word- and morpheme-based trigram language models [23]. Latent semantic analysis (LSA) is promoted by many researchers, such as Froud, Lachkar, and Ouatik (2012), who offer an LSA method that uses a variety of distance functions and similarity measures to determine the similarities between Arabic words. Their study compares the results for the use of the model with and without stemming. It was found that stemming affects the obtained results negatively when using the LSA model [25]. The same authors also used their system to produce new results for their previous experiment by comparing stemming and light stemming. The results showed that the light stemming approach out-performed the stemming approach because the latter affects the meanings of words [26]. In the medical domain, the LSA method has been used to predict protein-protein interaction, based on the Arabic semantic analysis model. This method was used to help the researchers understand how and why two proteins interact because protein pairs may interact if they contain similar or related Arabic words. This new method was compared with two other successful methods – namely, PPI–PS and PIPE, and higher accuracy was achieved with the new methods. This research gives insight, there-fore, into the importance of semantic analysis, as this method achieved more accurate results than other successful methods [27].

3. The methodology

A novel approach to improve the performance measures of informal Arabic language sentiment analysis is proposed to analyze the semantics and sentiment of user-generated text at the word and sentence level. We automatically capture patterns of words of similar contextual semantics and sentiment in tweets. The proposed approach does not depend on the syntactic structure of tweets; instead, it extracts patterns from the contextual semantic and sentiment similarities among words in a given tweet corpus. Contextual semantics are based on the proposition that meaning can be extracted from words' co-occurrences. We evaluate our approach by comparing the results of our approach with the results of the classic bigram and LSA approach. Figure1 illustrates the semantic sentiment analysis model for informal Arabic. An overview of the framework's four stages, as depicted in Figure 1, is presented in this section. The four stages of our framework are as follows:

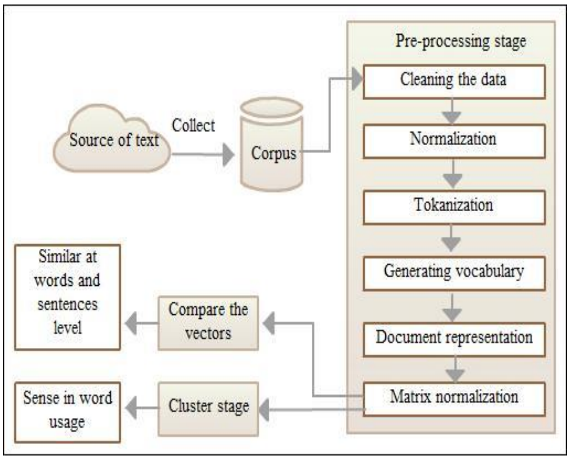

Figure 1. Framework for unsupervised clustering methodology.

3.1 *Dataset collection*

As seen in Figure 1 the first step is document gathering, which is conducted in order to build a corpus. We had to collect our own specialized data (i.e., tweets generated in KSA). For this research the Twitter mircoblog is one of the best resources for collecting our dataset. To collect the Arabic tweets, we used Twitter's stream API in order to avoid the problems of bias and excessive time consumption that can occur when collecting the data manually. The corpus contained 4,425,003 tweets that was saved in a database.

The data collection began on July 7, 2014. The duration of the data collection period coincided with the following events: the month of Ramadan, the FIFA World Cup, and Eid al-Fitr.

3.2 *Pre-processing*

The preprocessing stage is very important in achieving good results from text mining. In context of big data, it can also be seen as a preventive measure to handle the curse of dimensionality. Thus we created our own text preprocessing scheme to deal with informal Arabic language (i.e., Saudi dialect). The text preprocessing stage contains the following four steps:

1. *Cleaning the dataset*

   The cleaning process is used to remove all of the following cases:

- separate any non-Arabic word followed by Arabic word by single space, for example,
  Noorهلا -> Noor هلا
- separate any Arabic word followed by non-Arabic word by single space, for example,
  ارسلemail -> ارسل email
- replace all URLs with the symbol URL
- replace all emails with the symbol EMAIL
- replace all time formats with the symbol TIME
- replace all date formats with the symbol DATE
- replace all numbers with the symbol NUMBER
- remove repeated characters, for example, nooo-> noo
- remove repeated sequences no no no no no -> no no
- Separate symbol sequences, for example, ?!! -> ? ! !

We used this process for cleaning in order to reduce the corpus size and noise, while also ensuring that context of the tweet remains unchanged.

2. *Normalization*

   The normalization process is manipulating the text to produce consistent form, by converting all the various forms of a word to a common form. Table.1 shows the all normalization cases that we handled in our experiment.

Table 1 normalization cases

Rule	Example
Tashkeel	المؤمنين<-الْمؤْمِنينَ
Tatweel	الله<-اللـــــه
Alef	آor أor إ->ا
Heh	ەor ة-> ه

3.3 *Tokenization*

The tokenization process was performed for each tweet in order to divide the tweet into multiple tokens based on whitespace characters. The corpus was divided into 1,383,012 unique words.

3.4 *Generating vocabulary*

This process was used to build a list of vocabulary words that used the list of pairs (i.e., the word and its counts); the word order was arranged alphabetically. This resulted in 1,383,012 unique words. Then, to avoid out of memory problem, we reduced the vocabulary size to 13,696 words by deleting the words that appeared fewer than 400 times in the corpus, which equals 84% of the corpus. The computational and storage resources largely determined the frequency limit.

3.5 *Document representation*

In this step, these numerical data were transformed into vectors. The Bigram matrix was used to implement this task. The bigram matrix only contains numerical data. The Bigram matrix denoted by $X_{13696\times13696}$ has size of 13696 ×13696. Each entry in the matrix represents the frequency (i.e., how many times w_j came before w_i in the corpus). Figure 2 illustrates the process. The matrix contains the co-occurrence frequency for the words before and after; if we take sequence w_2 w_1, then word w_2 came before w_1, and if we take sequence w_1 w_2, then word w_1 came before w_2. In other words, w_2 came after w_1.

While the matrix is square, if we take the transpose of X (i.e.,X^T)we will be able to determine how many times w_j came after w_i. Then, concatenate the two matrixes together to make a new matrix and to make each vector contain the before and after frequency value. The new matrix is X = $[XX^T]$. The new size is n×2m, where n = 13696 and m = 13696.

$$
\begin{array}{cc}
 & \text{1}^{\text{st}}\text{ preceding word} \\
n \text{ Words} & \begin{pmatrix}
a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\
\vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\
a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{in} \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
a_{n,1} & a_{n,2} & \cdots & a_{n,j} & \cdots & a_{n,n}
\end{pmatrix}
\end{array}
$$

Figure 2 .Bigram matrix.

3.6 Normalization

The normalization helps prevent attributes with large ranges from outweighing attributes with smaller ranges (Jonker, Petkovic’, Lin, & Selcuk Candan, 2004). The bigram matrix for any given training corpus is sparse; most of the elements in the matrix are zero. This task of reassessing some of the zero value and assigning them non-zero values is called smoothing. We add one to all the counts in the matrix called X. This algorithm is called add-one smoothing (Jurafsky & James,2000).

Then, we use the column wise method to normalize the columns in matrix X by summing the elements in each column, i.e. $\sum_{i=1}^{n} a_{ij}$, where j is the column number. Then divide each element in the matrix with the perspective column sum, i.e. $a_{ij}/\sum_{i=1}^{n} a_{ij}$. (Novak & Mammone, 2001) Then, the based 2 log is calculated for all elements in the matrix X, to make the data more normally distributed (Zhai, 2008). Then, the z-score for all elements X is computed by subtracting the mean and dividing by the standard deviation. This should first by apply in the columns level and then the rows level.

$$x_{norm} = \frac{x - \mu}{\sigma}$$

Where x is referring to the score, μ refers to mean and σ refers to standard deviation.

3.7 Clustering Stage

After normalizing the numerical data, the words dimensions were reduced by applied K-means algorithm to categorize the words by setting k = 200. After many experiments we arrived at k=200 as the best result for word clustering to capture patterns of words of similar contextual semantics and sentiment in tweets (see section 4 to see the experiment result). Figure 3 illustrates part of bigram matrix after normalization

	A	B	C	D	E	F	G
1	-2499.5	-2422.1	-777.85	-2387.2	-2324.3	-4497.8	2174.9
2	-2929.8	-2968.2	-1123.1	-3114.6	-2736.6	-3829.5	1603.2
3	-2671.7	-2496	-148.86	-2671.9	-2359.9	-2910.8	2380.4
4	-926.15	-928.25	-964.55	-827.53	-406.08	-1382.5	-538.74
5	164.68	17.282	1158.1	-282.03	1016.8	414.08	595.85
6	-286.54	-447.94	-23.38	-627.81	-46.362	-98.742	-208.17
7	-351.95	-381.5	-9.7788	-472.51	327.51	-478.33	-98.13
8	-1392.5	-1417.1	-1676.2	-1273	-741.05	-2205.5	-462.82
9	-434.74	-580.88	-182.7	-566.88	210.84	-618.31	38.445
10	198.17	-21.362	811.79	-229.82	641.53	98.06	562.57
11	-582.16	-523.38	373.11	-743.49	-4.1742	-632.66	629.31
12	-14.064	-363.34	1537.8	-406.97	214.62	-336.83	1501.8
13	429.52	234.68	1233.8	170.34	1031.3	505.72	403.42
14	-1540.7	-1265.5	-1635.9	-1305.6	-1293.7	-1182.3	-1045.2

Figure 3 .Screenshot of bigram matrix after normalization

To find the similar contextual semantics and sentiment for the sentence level, we calculated the average of the words’ vectors that appeared in the sentence in order to get a new vector for the sentence. If we have sentence Si = $\{w_1, w_2, \ldots, w_n\}$.

$$\text{Sentence vector} = \frac{\sum_{i=1}^{n} Vw_i}{\text{Total number of words}}$$

Where V_{wi} denote to the value of the words’ vector in the sentence S_i. For example, if we have a tweet: “I love Mac products”. The vector of each word:

”I”:[1,0,3],”love”:[1,1,5],”Mac”:[2, 0,3], ”products”: [0, 0, 2]

The sum = [4, 1, 13]

Sentence vector = [4,1, 13]/ 4 = [1, 0.20, 3.25]

One challenge in clustering of short text (e.g., tweets) is that exact keyword matching may not work well (Aggarwal & Zhai, 2012).This research overcomes this challenge and extracts patterns automatically of words of similar contextual semantics and sentiment in tweets.

3.8 The Model Validation

This stage evaluates the model by comparing the model results with the results of the bigram model and LSA model. The bigram model used the same vocabulary size to build the matrix and also used the same normalization process. The bigram matrix denoted as matrix V with size 13696×13696. The LSA model used feature extraction TF-IDF, and set the SVD rank to used feature extraction TF-IDF, and set the SVD(singular value decomposition) rank to K = 100. We try to set the K = 200, but hardware limitation did not permit to perform experiments with this setting.

4. The experiments result

In this set of experiments, the tailored bigram model [XX^T] was used. The words dimensions were reduced by utilizing K-means clustering to analyze the semantics and sentiments of user-generated text at word and sentence levels. The proposed method was then compared with the bigram [X] and LSA models. Overall, three types of experiments were conducted: two at the word level and one at the sentence level.

The novel approach proposed here does not depend on the syntactic structure of the tweets; rather, it extracts semantic and sentimental patterns from a given corpus of tweets.

Experiment A: Finding similarities between words

- **Objective**

The aim was to analyze the semantics and sentiments of tweets at the word level by automatically capturing words with similar semantics and sentiments.

- **Method**

The unlabeled dataset contained 4,425,003 tweets. A vocabulary of 13,696 tokens was generated and used to create bigram matrix denoted by X, with size = 13,696×27,392. The normalization process was then used. Then tailored bigram matrix was used to discover the most similar words to the query word by comparing between words' vector values in the matrix XX^T. If words have similar vectors in a matrix, then they tend to have some relatedness. The similarity/relatedness was found by comparing between the vectors (each vector contains 27,392 features) we found the similar words to the query word by comparing which vectors are very closed to the query vectors by using square Euclidean distance function (this matrix was very huge and attempt to open it resulted in "out of memory" problem).

The model arranges all the vocabulary (similar words) in descending order based on similarity with query word (from word that has most highest similar to lower similar word). We select only the most 10 similar words to make the comparison between models more easy and to make it more clear for reader.

The model was also tested by extracting some sentiment words. The proposed model revealed that words indicative of sentiment tend to have high similarity with words of the same sentiment polarity. The sentiment words do not have different context, the proposed model extract the words that have similarity in polarity or related to query word. If the query word is positive, the model extracts similar/related positive words.

- **Results**

The results, which were selected and analyzed at random, are shown in tables 2, 3, 4, and 5.

The word sense is set of different meaning of the query word. In the tables 2 and 3 the number refers to how many meanings of word were discovered with respect to usage in text based on the context. In tables 4 and 5 results for sentiment are given. The sentiment words do not have different context, the proposed model extract the words that are similar in polarity or related to query word.

Table 2 List of the ten words most similar to هلال/helal

Word	LSA model	Bigram Model	Proposed tailored bigram model
هلال/ *helal*	بقدوم/ Advent شهر /Month تهنئه /Congratulates لشهر /For a month سدير /*Sadir* قدوم /Advent رمضان /Ramadan تعذر /Cannot مغرد/Twitter's	بقدوم/ Advent ياهلال /*Ya-helal* ست /Six بشهر /In a month عتقاء / Redeems تستقبلون / Greet شهر /Month لشهر /For a month	الهلال /*Al-helal* ياهلال /*Ya-helal* شهر /Month بعثه /Expedition ست /Six زعيم /Leader اتحاد /*Etihad* الغائبين /Absentee

	user المبارك /Blessed	بحلول /Advent لاخر / For the last	رشيد /Reashed سعد /Sad
Word sense	1	1	3

Table 3 List of ten words most similar to هدف/ goal

Word	LSA model	Bigram Model	Proposed tailored bigram model
هدف/ Goal	البوسنه /Bosnia	قوول /Gooal	قوول /Gooal
	هدفين /Two goals	بهدف /By goal or with aim	اهداف /Goals or the aims
	للمنتخب /To the team	اهداف /Goals or the aims	الهدف /The goal or the aim
	المونديال /World Cup	جوول /Goooal	هدفين /Two goals or two aim
	نيمار /Neymar	لاعب /Player	لاعب /Player
	التعادل /Equalizer	تسديده /Win	بهدف /By goal or with aim
	الارجنتين /Argentina	تسديده/Score	جوول /Goooal
	رونالدو /Ronaldo	مباراه/Match	التعادل /Equal
	مباراه /Match	مدرب /Coach	فوز /Win
	كلوزه /Klose	منتخب /Team	مباراه /Match
Word sense	1	2	2

Table 4 List of ten words most similar to سعاده/happiness

Word	LSA model	Bigram Model	Proposed tailored bigram model
سعاده/ Happi-ness	راحه/Comfort	وسعاده/And happi-ness	راحه/Comfort
	ابديه/Eternal	سعاده /Happily	وسعاده/And happiness
	جنه/Heaven	راحه/Comfort	بسعاده/Happily
	امي/My mother	سعيدةHappy	فرح/Joy
	وفرح/And joy	راحه/Comfort	فرحه/Joy
	اكتفاء/Satisfaction	طمانينه/Tranquility	جنه/Paradise
	اسعدها/Her happy	وبركه/And bless-ing	طمانينه/Tranquility
	♡ ~/Heart symbol	معالي /Excellency	حياه/Lifetime
	امنيه/Wish	عافيه/Good health	خير/Good
	قلبها/Her heart	جنه/Heaven	رضا/Satisfaction

Table 5 List of ten words most similar to حزن/sadness

Word	LSA model	Bigram Model	Proposed tailored bigram model
حزن/ sad-ness	قلق /Concern	وحزن / And Sadness	وحزن / And Sadness
	خذلان /Betrayal	اشتياق /Longing	اشتياق /Longing
	يكتمل /Complete	الم /Pain	حنين /Nostalgia
	فقر /Poverty	فرح/Joy	الم /Pain
	خيبه /Disappointment	جرح /Wrench	شوق /Longing
	تسكن /Live	ضيق /Restless	فرح/Joy
	بلا /Without	وجع /Wrench	وجع /Wrench
	محال /Impossible	شوق /Longing	حب /Love
	خبل /Dementia	الحزن /Sadness	ضيق /Restless
	دنيا /World	النقاء/purity	جرح /Wrench

- **Discussion**

Tables 2 to 5 show the words most similar/related to the given query word by using the proposed tailored bigram, bigram and LSA models. All these methods capture broad semantic and sentiment relatedness. Based on human evaluation of the experiments results, the tailored bigram model seems to perform better than the LSA model because the proposed model captures more different semantic and sentimental related patterns from a given corpus of tweets.

As can be seen from Table 2, the proposed model shows how the word هلال/helal can have different meanings according to the context; it can mean "crescent," it can be the name of a Saudi football team, or it can be the name of a person. The word 'هلال'/helal has the meaning "crescent" and is similar to the word 'شهر'/month and the word 'ست'/six, which could be denoted as "number" or "date." Furthermore, helal, in its meaning as the name of a Saudi football team, is similar to the word زعيم/leader, which is the nickname of the team. The word اتحاد/Etihad is also the name of a Saudi football team. In its meaning as a person's name, helal is similar to رشيد/Reashed and سعد/Sad, which are also people's names. The LSA model only gives one semantic context, which is the word helal meaning only "crescent."

In Table 3, the proposed model shows how the word هدف/goal can have different meanings according to the context; it can mean "score a goal," and it can mean "aim or target." Also, the proposed model extracted some informal words, such as قوول/gooal and جوول/goooal, which are similar to the word هدف/goal. The LSA model only gives one semantic context, which is 'score a goal."

The model was also tested by extracting some sentiment words, as shown in tables 4 and 5. The proposed model revealed that words indicative of sentiment tend to have high similarity with words of the same sentiment.

Similarly, in Table 5, the proposed model presents the words فرح/joy and

حب/love as being similar to the word حزن/sad—again, all of them have similar words that come after and before then in a sentence. The LSA extracted four words: يكتمل/Complete, تسكن/Live, بلا/Without and دنيا/World were not similar to word sad.

Experiment B: Testing the clustering results

- **Objective**

The aim was to analyze the semantics of tweets at the word level by automatically capturing patterns of words with similar contextual semantics by using the proposed model (i.e., tailored bigram) which was found to have the highest performance level in the previous experiment.

- **Method**

The train set contained 4,425,003 tweets. A vocabulary of 13,696 tokens was generated and used to create a bigram matrix denoted by X with size = 13696×27392. K-means clustering was then used to categorize the words into k = 100; 200, and 300 clusters. then, made a comparison between k values and we arrived at k = 200 as the best result for word clustering to capture patterns of words of similar contextual semantic and sentiment in tweets. Each row was called a vector, and each column was called a dimension (each representing a different semantic feature).

- **Results**

The results are shown in Tables 6, 7 and 8.

Table 6 Extracted patterns for the semantic word/هلال helal

Word	Proposed tailored bigram model			
هلال / Helal	Dim 1	Dim 2	Dim 3	Dim 4
	بقدوم /Coming	عبدالله /Abdullah	الغامدي / Al-Ghamdi	تشكيله/ Formation
	بحلول /Advent	سعد /Sad	الحربي / Al-Harbie	لاعبي /Players
	بشهر/At month	عبدالرحمن/ Abdulrahman	القحطاني / Al-Qahtani	جماهير /Masses
	بمناسبه /Occasion	فهد /Fahd	الشمري / AlShammari	مباراه /Match
	بالعيد /Eid	خالد /Khalid	الدوسري / Al-Dosari	فوز /Win
	بقرب /Near	بندر /Bender	العنزي/Al-Anzi	مرمى /Goal
	برمضان /Ramadan	ناصر /Nasser	بن /Son of	يشجع /Encourages
	قدوم /Coming	فيصل /Faisal	الشهراني / Al-Shahrani	مدرب/Coach

Table 7 Extracted patterns for semantic word/هدف goal

Word	Proposed tailored bigram model			
هدف/ Goal	Dim 1	Dim 2	Dim 3	Dim 4
	تشكيله/ Formation	نيمار/ Neymar	قرار /Decision	بلجيكا/ Belgium
	لاعبي/ Players	سواريز/ Suarez	طلب/ Request	هولندا/ Netherlands
	جماهير/ Masses	ميسي/Messi	حساب/ Account	الارجنتين/ Argentina
	مباراه/Match	سانشيز/ Sanchez	موضوع/ Topic	المانيا/ Germany
	فوز/Win	كوستا/Costa	خبر/News	تشيلي/Chile
	مرمى/Goal	بنزيما/ Benzema	موقع/ Position	فرنسا/France
	يشجع/ Encourages	روني/ Ronnie	اسم/Name	غانا/Ghana
	مدرب/Coach	رونالدو/ Ronaldo	بيان/ Statement	كولومبيا/ Columbia

Table 8 Extracted patterns for the semantic word خطير/Dangerous

Word	Proposed tailored bigram model			
خطير/ Dangerous	Dim 1	Dim 2	Dim 3	Dim 4
	رائع/ Wonderful	هلالي /Hilali	محظوظ /Lucky	موجع /Painful
	جيد/Good	برازيلي /Brazilian	محترم /Respected	مؤلم /Painful
	مميز/Special	نصراوي /Nasraoui	مبسوط /Happy	المؤلم /Painful
	ممتع /Fun	سعودي/ Saudi	اناني /Selfish	يوجع /Pain
	سيء /Bad	اهلاوي /Ahlawy	غلطان /mistaken	يؤلمني /Pains me
	جميل/ Beautiful	عربي/ Arabian	كذاب /Liar	يوجعني/ Pains me
	ممتاز/ excellent	مدريدي /Madrida	مظلوم /oppressed	يوجعك /Pain
	كبير /Big	مصري /Egyptian	مجنون /crazy	قاسي /tough

Tables 6 to 8 give the results for the four most common dimensions for the given query words using the proposed model after reduction of the dimensions to 200 features. All these query words captured broad contextual semantic similarities. K-means clustering was used to determine which words belonged to each cluster.

Table 6 presents the top four semantic features (or dimensions) for the word هلال/helal. The first dimension indicates the results obtained from mining the meaning “crescent.” The second dimension is related to the word’s meaning as a person’s name, and the third dimension indicates the word's meaning as a family name. The fourth dimension is connected with the meaning of helal as the name of a Saudi football team.

Table 7 presents the key semantic features of the word goal. The first dimension indicates the results obtained from mining the meaning "score a goal." In the second dimension, the word indicates the names of football players, which are also in the sport domain. In the third dimension, the word gives the meaning of the "aim or target". The fourth dimension connects the word with its meaning in relation to the names of countries. The words and phrases goal, player, team name, and country name were all found before the word سجل /score, explaining why these words appeared in these particular dimensions.

In table 8 the four most common semantic features for the word dangerous are presented. The first dimension refers to the word's meaning as "wonderful" (i.e., positive). The second dimension is not related to the word dangerous. The third dimension indicates the first three results gave the word dangerous meaning wonderful (i.e., positive), and the last five results gave the word dangerous meaning bad (i.e., negative). The forth dimensions refers to the word's meaning as dangerous (i.e., negative).

The proposed model categorizes words together that have similar semantic features, automatically capturing the contextual patterns in tweets. If a word has multiple contextual meanings, the model uncovers these meanings, adding each word to the relevant clusters.

<u>Experiment C: Finding similarities at sentence level</u>

- **Objective**

The aim is to analyze the semantics and sentiments of tweets at the sentence level by automatically capturing sentences with similar semantics and sentiments. The words هلال/helal, هدف>/goal, and sentiment word خطير/Dangerous were tested in the results for the proposed model at the sentence level.

- **Method**

First, all the sentences in the database that contained the query words were extracted. The vector average of each sentence was then calculated using the bigram matrix X after reducing the dimensions as input in equation (2), thus giving a new matrix, V si×200. The dimensions were reduced because the bigram matrix used in calculating matrix V caused a memory problem with the computer. The reduction eliminated the problem

The five sentences that were most similar to each query sentence were tested by comparing the sentence vectors using matrix V.

- **Results**

The results are shown in tables 9, 10, 11, 12, 13, 14 and

Table 9 The five most similar sentences that contain the word هدف/goal, where the word هدف means "goal"

Original Sentence	Most similar sentences
نيمار خامس لاعب يسجل هدف من خارج منطقه الجزاء في مباراه افتتاحيه كاس العالم "Neymar was the fifth player to score a goal from outside the penalty area in the opening World Cup match"	هدف جيميس جوول يامعلمهم جيميس "James goal, hey teacher James"
	قوول *الهدف* الاول لكولمبياضربه جزاءالبرازيل NUMBER كولمبيا NUMBER هدف الاول رودريغيز URL "Saay the first Colombian goal, a penalty. Brazil NUMBER Columbia NUMBER, first goal from Rodriguez URL
	الهدف الاول للمنتخب الكولومبي من ركله جزاء #البرازيل كولومبيا "Colombia's first goal for the team from the penalty spot # BrazilColombia"
	الجزائر بالغت في الدفاع في الشوط الثاني فاستغل @albakertv بلجيكا هذا التراجع وقتل امال الجزائر بهدف قاتل استراليا NUMBER هولندا NUMBER. @albakertv Algeria exaggerated the defense in the second half, Belgium took advantage of this decline and killed the hopes of Algeria by a murderous goal Australia NUMBER Netherlands NUMBER.
	هدف يشعلل المباراه *(: "The goal ignites the match :) *"

Table 10 The five most similar sentences that contain the word هدف/goal, where the word هدف means “aim”

Original Sentence	Most similar sentences
عندما يكون لديك هدف لا تسلط تركيزك على المستقبل “ الان “وتخسر ،، سلط الضوء “ الان “على وهي حتما ستصلك الى ما تريد” “When you have a goal, do not fix your focus on the future and lose the ‘now’; highlighted the ‘now’ and, inevitably, you will receive what you want.”	لا يوجد شخص ولد كبيرا .. ولا يوجد مشهور لم يبدا صغيرا .. ولا يوجد هدف تحقق الا وكان حلما وليداالبدايه من الصفر ليست عيبا “No person is born an adult. No famous person does not start small. There is no goal archived, until there was a dream that born from scratch, this is not a defect”
	@alfofoNUMBER : لا تعيش بلا امال ولا هدف , ولا تقول ان الزمن دايم بخيلاجمل الاشياء تلقاها صدف .. والفرح دايم ورى الصبر الجميل !! “@alfofoNUMBER: do not live without hope nor objective, and do not say that time is always miserly; make things beautiful, received coincidence, and find joy always behind beautiful patience!!”
	حدد هدفك واستعن بالله ببساطه كم تريد ان تختم كتاب الله في رمضان #الاستعداد لرمضان #رمضان “Simply how many times you want to seal the Book of Allah in Ramadan? Determine your goal and seek the help of God #prepare for Ramadan #Ramadan.”
	@ zlfay qNUMBERmeNUMBER@ aNUMBERm@ abnazulfi@ ولك الشكر ومن تفاعل وهدفنا جميعا المصلحه العامه حفظ الله الوطن والقائمين عليه متمنين ان تحل هذه المشكله “@QNUMBERmeNUMBER @aNUMBERm @abnazulfi @zlfay and thanks to you also for how interact, and our aim of all is the public interest. God Save the homeland and those who support it and wish that interaction would solve this problem.”
	صوره مؤثره شاب سوري يطعم طفله من تحت باب منزلها بهدف طمانتها والتخفيف عنها An image of an inuential young Syrian feeding a child from under the door of her home in order to reassure her and to comfort her

Table 11 The five most similar sentences that contain word هلال/helal, where the word هلال means “the name of a Saudi football team”

Original Sentence	Most similar sentences
مبروك بطوله الليموزين يا هلال URL “Congratulations Championship Limousine O Helal URL”	يسالوني من عقب الهلال تحب !؟ # الهلال يابعدي مابعد ه احد “Ask me after the al-helal love!? # al-helal my love, no one came after it.”
	تتواجد عدد من الجماهير العمانيه بمقر نادي # الهلال جاءت لتؤازر الفريق الازرق امام السد .URL “Number of Omani fans present at the Club #al-helal came to co-operate the blue team in front of al-Sad. URL.”
	“@bluegoldNUMBER: قبل NUMBER سنوات فتح الغرافه الملعب بالكامل لجمهور الهلال .. لانها اخلاق الكبيرولا عزاء لابو خمسه .. !! URL “@ bluegoldNUMBER: NUMBER years before Gharafa opens the entire stadium to the al-Helal audience. because this is manners of the greater no consolation to Abu five .. !! URL.”
	لو الهلال حقق اسيا وش بتسويفي حال اقام الاتحاد # الاسيوي بطوله اسيا بمشاركه الهلال وحده فقط ، اقول ربما يحقق البطوله ..... اقول ربما . “If al-Helal achieved the champion of Asia, what you will do in case if the AFC established Asian Cup with the participation of al-Helal alone only, say perhaps he will achieve tournament ... .. I side probably.”
	لما تكون نادي كبير والرقم الاول جماهيرا ، فانت تحتاج الى مركز اعلامي قوي جدا !! وما يحدث في نادي # الهلال هو العكس تماما !@ a bin mosaad “When you become a big club and the first number to be a mass, you need a media center that is very strong!! What happens in Club #al-Helal is quite the opposite! @abinmosaad”

Table 12 The five most similar sentences that contain the word هلال/helal, where the word هلال means “the name of a person”

Original Sentence	Most similar sentences
وابغ من العيش ما تسر به ان عذل الناس فيه او عذروابو هلال العسكري “I want from the living what pleased, if Conquer people with it Abu Hela al-Askarry.”	الشهاده لله ان نفتخر بالشباب الي مثل اخي @alialnimi ابو هلال سباق لفعل الخير جزاك الله خيرا . ونحن في الخدمه “@TheNaim testimony to God that we are proud to youth like my brother Abu Helal, he racing to do good things, God reward you. We are in the service any time.”
	بس ما يقدرون لانهم mbc الهلاليين يبون يقاطعون بيتابعون السعوديه الاولى “Alhlaliyn wants to boycott MBC but they cannot because they follow first Saudi channel.”
	تطرح شركه صله تذاكر مباراه # الهلال والعروبه اليوم NUMBER عصرا حتى NUMBER بالنادي من الساعه مساء ويتواصل بيع التذاكر غدا بالملعب باستاد الملك فهد “The Sela company raises the number of tickets for the #Al-Helal and Arabism match today from NUMBER pm until NUMBER evening and will continue selling tickets tomorrow at King Fahd Stadium”
	متى يلعب الهلال واشوف التايم لاين كله ازرق والجمهور يملي مدرجات الهلال ونشوف اللاعبين متى يجي اليوم اللي اكون مبسوطه كثير “When al-Helal playing and see the whole twitter favorite is blue and the al-Helal’s fans full the stadium, and see players When the day comes when I can be joyous.
	سبق اني ارسلت تغريده تخص ديانه الهلال مصدرها صحيفه سبورت الاكترونيه وبعد ان وضح ان ما كتب غير صحيح اترفع باخلاقي كمسلم نصراوي واعتذر “I already sent a tweet about al-Helal religion source electronic newspaper Sport and after it explained that what has been written is not true, as Muslim manners and Nasraoui, I apologized.”

Table 13 The five most similar sentences that contain the word هلال/helal, where the word هلال means “crescent”

Original Sentence	Most similar sentences
عاجل المشروع الاسلامي لرصد الاهله : تمت رؤيه هلال شهر شوال قبل قليل نهارا في السعوديه باستخدام تقنيه التصوير الفلكي #هلال شوال # رمضان “Urgent Islamic Crescents Observation Project: This sighting of the new moon of Shawwal shortly before daybreak in Saudi Arabia uses the technique astrophotography #crescentofShawwal #Ramadan.””	كل العالم ينتظر هلال الفطر (شوال) الا اهل غزه فانهم ينتظرون هلال النصرويقولون متى هوقل عسى ان يكون قريب “All the world is waiting for the crescent al-Fitr (Shawwal), but the people of Gaza, they are waiting for the crescent’s win and say when it Tell It may be that close.”
	الجهات الحكوميه @balsayegh@ arabicobama بالسعوديه تتبع تقويم ام القرئ وليس رؤيه الهلال “@Balsayegh @arabicobama government agencies in Saudi Arabia follow the calendar or Imm Alqri; they do not see the crescent.”
	الاتحاد سجل هدف وسنتر الهلال وتلعب الكور ه في نصف الملعب وبعد دقائق يلغى ا لحكم هدف الاتحادURL عرفتو كيف البطوله = NUMBER “Al-Etehaad scored and al-Helal’s center and play a ball in half pitch and after minutes the ruling canceled Al-Etehaad goal URL know now how the championship = NUMBER.”
	بامكانك تحديد حجم عقليه المشجع الهلالي من خلال حجم طاقيته - طاقيه صغيره - يقولك " السابعه تقترب "- طاقيه كبيره - يقولك " الهلال ملكي ".

	"You can determine the size of the mentality encouraging Hilali through the cap size—small cap tells you 'seven o'clock approaching'; large cap tells you 'Royal al-Helal.'"
	@eeNUMBERqwe انا اموت ولا اغير الهلال "@eeNUMBERqwe If I die, it does not change my al-Helal."

Table 14 The five most similar sentences that contain the word Dangerous, where the word Dangerous means "Dangerous".

Original Sentence	Most similar sentences
طرق NUMBER لشرب الشاى تجعله خطير جدا على صحتك URL NUMBER ways to drink tea to make it very dangerous to your health URL	جالس اتابع مباراه العين والنصر اعاده اجانب العين مافيهم الا جيان الخطير اما كيمبو والسولفاكي والكوري مستواهم عادي الحمد الله I am following the rematch between Al-Aean and al-Nasser, at al-Aean, the foreign player Jian is dangerous but Kimpo and Alsolgaki and Korea are normal
	دنبلي خطير جداوسريع Denbla very dangerous and fast
	فتوى ان من اجاز الاغاني مجاهرلاتجوز امامته رايي انها زله خطيره The fatwa of authorized songs boldness may not be Imamth my opinion it slip serious
	@ soumahran ايه يابنتي جيتي المستندات الخطيره دي منين hi from where you got this Serious Documents
	عباره خطيره تؤدي الى الشرك وهي ان تقول [بكره يحلها الف حلال] والصح [يحلها الله سبحانه] فقول الف حلال تعني " ان هناك الف رب " لااله الا الله Dangerous word lead to polytheism which is to say [tomorrow will solve it thousand solver] the correct say [solved by God] To say a thousand solver means that there are a thousand Lord "to God but God."

Table 15 The five most similar sentences that contain the word Dangerous, where the word Dangerous means "Wonderful".

Original Sentence	Most similar sentences
@ fahdalruqi خطير و الله ابو عمر و شاعر بعد @ fahdalruqi God you are wonderful Abu Omar, and you also poet	@ tntn NUMBER تابعيها مرهه خطيره @ tntn NUMBER follow her she is wow
	اصابه نيمار الخطيره في مباراه البرازيل وكولومبيا .. كسر في فقره الظهر خطيره جدا URL Neymar had serious injury in the match between Brazil and Colombia .. a broken vertebra back very serious URL
	انه امر خطير جدا استقبال النكت باستمرارخطير على تكوين ابناؤنا وبناتنا It's too dangerous receiver jokes constantly risk of the formation of our sons and daughters
	بانزيما خطير والله خطير قوول Benzema goal is grave and serious
	@ noda NUMBER الصوره خطيره ✌ @ Noda NUMBER Image is serious ✌

- **Discussion**

Tables 9 to 15 give the five most similar sentences to the given query sentences using the word representations generated by the proposed model. All these vectors capture broad semantic and indirect sentiment similarities.

Table 9 gives the five sentences containing the word goal that are most similar to the query sentence where the word goal means "scored a goal."

Table 10 gives the five sentences containing the word goal that are most similar to the query sentence, where the word goal means "aim" or "target." The proposed model extracts similar semantic contextual sentences that contain the word goal where it means "aim" or "target." The results in table 8 and 9 show the two different semantic contexts for the word goal at the sentence level.

Table 11 presents the five sentences containing helal that are most similar to the query sentence, where the word helal means "the name of a Saudi football team." All the similar sentences contain helal where it denotes the name of a Saudi football team.

Table 12 gives the sentences where helal means "the name of a person." The proposed model only extracts the one sentence where helal means "a person's name" in the sentence context. The other similar sentences refer to helal as the name of a Saudi football team and are not similar to the query sentence.

Table 13 presents the sentences where helal means "crescent." Here, the model has only extracted two sentences where the word helal means "crescent;" the other three are examples where the word helal refers to the Saudi football team and are not similar to the query sentence. The results in table 11, 12, and 13 show the three different semantic contexts for the word helal at the sentence level.

Table 14 presents the sentences where Dangerous, where Dangerous means "Dangerous". The model has extracted the five sentences containing the word Dangerous that are similar to the query sentence, where Dangerous means "Dangerous". Table 15 presents the sentences where Dangerous means "Wonderful". Here, the model has only extracted gives the three sentences containing the word Dangerous that are similar to the query sentence, where Dangerous means "Wonderful". The results in table 14 and 15 show the two different semantic contexts for the word Dangerous at the sentence level.

Our proposed model has been used to analyze the semantics and sentiments of tweets at the sentence level, automatically capturing the patterns of sentences with similar contextual semantics and sentiments in tweets. According to the model results, the method needs to be developed further in order for more accurate results to be obtained.

5. Conclusion

The proposed tailored bigram model used unsupervised clustering at word and sentence level to allow semantic and sentiment categorization to take place. In the experiments, words and sentences in tweets with similar semantics and sentiments were automatically captured and grouped. The proposed model was then compared with the classic bigram and LSA models. Our proposed approach was not concerned with the syntactic structure of tweets, but with the extraction of patterns in semantics and sentiments from a particular tweet corpus.

With this methodology, a huge corpus was used, no annotation processing was utilized for labels, the word order within the tweets was considered, and no filtering process was used. The filtering was used only to "clean" the text, thus reducing the corpus size and the noise in the text. These steps were taken to ensure that the contexts of the tweets remained unchanged. Semantic dictionaries or lexicons were not used due to their limited coverage for informal Arabic. Based on our work, we conclude that although difficult to handle, big data can help in checking almost every type of possibility of similarity/ relatedness among words. Although due to availability of limited computational resources, we used some threshold to reduce the data, but were still were able to get good results. The manual evaluations of the results need to be automated for which Arabic semantic resources should be developed.

References

[1] L. Albraheem and H. S. Al-Khalifa, "Exploring the problems of sentiment analysis in informal Arabic," in *Proceedings of the 14th international conference on information integration and web-based applications and services*, 2012, pp. 415–418.

[2] A. Shoukry and A. Rafea, "Sentence-level Arabic sentiment analysis," in 2012 international conference on collaboration technologies and systems (CTS), 21-25 May 2012, 2012, pp. 546–550.

[3] B. Liu, "Sentiment analysis and opinion mining," Synth. Lect. Hum. Lang. Technol., vol. 5, no. 1, pp. 1–167, 2012.

[4] R. T. Khasawneh, H. A. Wahsheh, M. N. Al-Kabi, and I. M. Alsmadi, "Sentiment analysis of Arabic social media content: A comparative study," in 2013 international conference on Information Science and Technology (ICIST), 9-12 Dec. 2013, 2013, pp. 101–106.

[5] M. Rushdi-Saleh, M. T. Martín-Valdivia, L. A. Ureña-López, and J. M. Perea-Ortega, "OCA: Opinion corpus for Arabic," J. Am. Soc. Inf. Sci. Technol., vol. 62, no. 10, pp. pp. 2045–2054, 2011.

[6] H. Saif, Y. He, M. Fernandez, and H. Alani, "Semantic patterns for sentiment analysis of Twitter," *Semant. Web–ISWC 2014*, vol. (Vol. 8797, pp. pp. 324–340), 2014.

[7] C. C. Aggarwal and C. Zhai, Mining text data. Berlin & Heidelberg, Germany: Springer, 2012.

[8] C. Zhai, "Statistical language models for information retrieval," Synth. Lect. Hum. Lang. Technol., vol. 1, no. 1, pp. 1–141, 2008.

[9] J. Lin and C. Dyer, "Data-intensive text processing with MapReduce," *Synth. Lect. Hum. Lang. Technol.*, vol. 3, no. 1, pp. 1–177, 2010.

[10] M. Moussa, M. W. Fakhr, and K. Darwish, "Statistical denormalization for Arabic text," in Empirical Methods in Natural Language Processing, 2012, vol. 228, pp. 228–232.

[11] W. Naptali, M. Tsuchiya, and S. Nakagawa, "Word co-occurrence matrix and context dependent class in LSA based language model for speech recognition," Int. J. Comput., vol. 3, no. 1, pp. 1–11, 2009.

[12] D. Laniado and P. Mika, "Making sense of Twitter," in The Semantic Web–ISWC 2010, vol. 6496, P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, and B. Glimm, Eds. Berlin & Heidelberg, Germany: Springer, 2010, pp. 470–485.

[13] R. M. Duwairi, R. Marji, N. Sha'ban, and S. Rushaidat, "Sentiment Analysis in Arabic tweets," in 5th international conference on information and communication systems (ICICS), 1-3 April 2014, 2014, pp. 1–6.

[14] A. E.-D. A. Hamouda and F. E. El-Taher, "Sentiment analyzer for Arabic Comments System," Int. J. Adv. Comput. Sci. Appl., vol. 4, no. 3, pp. 99–103, 2013.

[15] M. M. Itani, L. Hamandi, R. N. Zantout, and I. Elkabani, "Classifying sentiment in Arabic social networks: Naive Search versus Naive Bayes," in 2012 2nd international

conference on advances in computational tools for engineering applications (ACTEA), 12-15 Dec. 2012, 2012, pp. 192–197.

[16] S. R. El-Beltagy and A. Ali, "Open issues in the sentiment analysis of arabic social media: A case study," in 2013 9th international conference on innovations in Information Technology (IIT), 17-19 March 2013, 2013, pp. 215–220.

[17] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. Al-Ayyoub, "Arabic sentiment analysis: Lexicon-based and corpus-based," in 2013 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT), 3-5 Dec. 2013, 2013, pp. 1–6.

[18] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," Comput. Linguist., vol. 37, no. 2, pp. 267–307, 2011.

[19] H. Saif, Y. He, and H. Alani, "Semantic sentiment analysis of Twitter," in The Semantic Web–ISWC 2012, vol. 7649, P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, J. Euzenat, M. Hauswirth, J. X. Parreira, J. Hendler, G. Schreiber, A. Bernstein, and E. Blomqvist, Eds. Berlin & Heidelberg, Germany: Springer, 2012, pp. 508–524.

[20] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in Proceedings of the 49th annual meeting of the Association for Computational Linguistics: human language technologies, 2011, vol. 1, pp. 142–150.

[21] H. Froud, A. Lachkar, and S. A. Ouatik, "Arabic text summarization based on latent semantic analysis to enhance Arabic documents clustering," Int. J. Data Min. Knowl. Manag. Process, vol. 3, no. 1, pp. 79–95, 2013.

[22] H. M. Alghamdi, A. Selamat, and N. S. A. Karim, "Arabic web pages clustering and annotation using semantic class features," J. King Saud Univ. Inf. Sci., vol. 26, no. 4, pp. 388–397, 2014.

[23] R. Sarikaya, M. Afify, Y. Deng, H. Erdogan, and Y. Gao, "Joint morphological-lexical language modeling for processing morphologically rich languages with application to dialectal Arabic," IEEE Trans. Audio. *Speech. Lang. Processing*, vol. 16, no. 7, pp. 1330–1339, 2008.

[24] A. E.-D. Mousa, R. Schluter, and H. Ney, "Investigations on the use of morpheme level features in language models for Arabic LVCSR," in *2012 IEEE international conference on acoustics, speech and signal Processing (ICASSP), 25-30 March 2012*, 2012, pp. 5021–5024.

[25] H. Froud, A. Lachkar, and S. A. Ouatik, "Stemming for Arabic words' similarity measures based on Latent Semantic Analysis model," in 2012 international conference on multimedia computing and systems (ICMCS), 10-12 May 2012, 2012, pp. 779–783.

[26] H. Froud, A. Lachkar, and S. A. Ouatik, "Stemming versus Light Stemming for measuring the simitilarity between Arabic Words with Latent Semantic Analysis model," in 2012 colloquium in Information Science and Technology (CIST), 22-24 Oct. 2012, 2012, pp. 69–73.

[27] N. M. Zaki, K. A. Alawar, A. A. Al Dhaheri, and S. Harous, "Protein-protein Interaction Prediction using Arabic semantic analysis," in 2013 9th international conference on innovations in Information Technology (IIT), 17-*19 March 2013*, 2013, pp. 243–247.