Sentiment Analysis of Iraqi Arabic Dialect on Facebook Based on Distributed Representations of Documents

ANWAR ALNAWAS, Department of Computer Engineering, Faculty of Technology, Gazi University, Turkey/Nasiriyah Technical Institute, Southern Technical University, Iraq
NURSAL ARICI, Department of Computer Engineering, Faculty of Technology, Gazi University, Turkey

Nowadays, social media is used by many people to express their opinions about a variety of topics. Opinion Mining or Sentiment Analysis techniques extract opinions from user generated contents. Over the years, a multitude of Sentiment Analysis studies has been done about the English language with deficiencies of research in all other languages. Unfortunately, Arabic is one of the languages that seems to lack substantial research, despite the rapid growth of its use on social media outlets. Furthermore, specific Arabic dialects should be studied, not just Modern Standard Arabic. In this paper, we experiment sentiments analysis of Iraqi Arabic dialect using word embedding. First, we made a large corpus from previous works to learn word representations. Second, we generated word embedding model by training corpus using Doc2Vec representations based on Paragraph and Distributed Memory Model of Paragraph Vectors (DM-PV) architecture. Lastly, the represented feature used for training four binary classifiers (Logistic Regression, Decision Tree, Support Vector Machine and Naive Bayes) to detect sentiment. We also experimented different values of parameters (window size, dimension and negative samples). In the light of the experiments, it can be concluded that our approach achieves a better performance for Logistic Regression and Support Vector Machine than the other classifiers.

CCS Concepts: • **Information systems → Sentiment analysis**;

Additional Key Words and Phrases: Doc2Vec, Iraqi Arabic Dialect, word embedding, sentiments analysis, facebook

ACM Reference format:
Anwar Alnawas and Nursal Arici. 2019. Sentiment Analysis of Iraqi Arabic Dialect on Facebook Based on Distributed Representations of Documents. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 18, 3, Article 20 (January 2019), 17 pages.
http://doi.org/10.1145/3278605

1 INTRODUCTION

Nowadays, Social networking sites have generated a tremendous amount of data. These data contain a high level of opinions and user generated emotions on specific topics [1]. Therefore it is necessary to analyze these feelings about specific topics [2]. Thus, Sentiment Analysis (denoted

Authors' addresses: A. Alnawas, Department of Computer Engineering, Faculty of Technology, Gazi University, 06500 Ankara, Turkey/Nasiriyah Technical Institute, Southern Technical University, Iraq; emails: anwaradnanmzher.alnawas@gazi.edu.tr, anwar.alnawas@stu.edu.iq; N. Arici, Department of Computer Engineering, Faculty of Technology, Gazi University, 06500 Ankara, Turkey; email: nursal@gazi.edu.tr.

2375-4699/2019/01-ART20 $15.00
http://doi.org/10.1145/3278605

as SA) has become a common task in this area, which aims to determine the polarity of written comments.

Hence, SA has gained attention in many areas (e.g. business analytics) because of its simplicity and efficiency. Many products are being developed based on users' opinions from consumer reviews about a given product to the reactions surrounding the political aspect [3, 4].

Many research on sentiment analysis targeted English language, but there have been recent efforts to expand the research on other languages like Arabic. Arabic SA is considered a complex problem because of the challenging features of Arabic language. In addition, it has a complex morphology, it is highly inflectional and derivational compared to other languages [5].

Arabic is one of the most Morphologically Rich Languages (MRL). This means manual feature extraction is a challenging and time consuming task. Optimal extraction of opinions depends on good representation of reviews. Bag-of-Words is used as traditional approach to represent documents where word is presented in fixed-length. The limitations of this approach are missing the order of the word in the context of the sentence and ignoring the grammatical structure and lexicon dependent [6].

Arabic is one of the Semitic languages and consists of three subcategories; Classic Arabic, Modern Standard Arabic (MSA), and Dialectical Arabic. Among these three types, the dialects of Arabic language are difficult compared with existing varieties.

Most of the previous studies on Arabic SA have considered either classic Arabic or modern standard Arabic with a clear weakness towards dialects [7, 8]. However, dialects at present are not used for spoken communication only, but expanded its use as a written language, especially in different social media environments [9]. This adds complexity to dealing with texts in social networks because they are written in an informal style and different spelling forms. When applying traditional sentiment analysis techniques, this requires a lot of pre-processing to extract features. Therefore, word embedding can alleviate this problem using it as a source of features extraction. Recently, word embedding has been used for MSA sentiment analysis. However, the experiments conducted on MSA with a clear weakness in dialects [10].

The purpose of this paper is to experiment the SA of Iraqi Arabic Dialect, which has not been done so far to the best of our knowledge, based on word embedding as features extracting technique. We will apply Doc2Vec with four of supervised machine learning algorithms Logistic Regression (denoted as LR), Decision Tree (denoted as DT), Support Vector Machine (denoted as SVM) and Naive Bayes (denoted as NB) and compare the performances.

The rest of this paper was organized as follows: Related works were explained in Section 2. The background was presented in Section 3 and the methodology of our work was described in Section 4. In Section 5, we described the experimental setup and datasets that were used. In Section 6, we investigated the results and impact of the word embedding parameters on the classification performance. In Section 7, we discussed our observations in this work and finally in Section 8, we concluded the paper and remarked key findings.

2 RELATED WORK

There are limited works used for word embedding extract features in Arabic SA. Dahou, et al. [11] used publicly available datasets from [12–14]. They focused on building a word embedding using 3.4 billion words. To evaluate the performance of the created word embedding, they used CNN to train the Arabic word embedding.

In the work of Altowayan and Tao [3] word embedding was used to extract features for sentiments analysis model. Their model consists of three steps; compile a large Arabic corpus, generate word vectors (embedding) and detect subjectivity and sentiment by training several binary classifiers. They used Word2Vec tool from [15] to represents words vector. Their approach achieved

a slightly better accuracy when compared with other methods in the literature that was based on hand-crafted features.

The work of Al-Azani and El-Alfy [9] compared more than one classifier on highly imbalanced SA of tweets datasets. They fed word embedding features using Syria Tweets dataset. In the work of Al-Sallab, et al. [16], they handled morphological complexity and the lack of opinion resources for Arabic using Recursive Auto Encoder (RAE) model. The limitations of RAE when dealing with Arabic as they mention: the treatment morphological complexity of Arabic, input features are more complete and comprehensive for the auto encoder, and semantic composition and express the overall meaning carry out by following the natural way constituents. To addresses those limitations, they proposed A Recursive Deep Learning Model for Opinion Mining in Arabic (AROMA). The other study of Baly, et al. [17] evaluates the Recursive Neural Tensor Networks (RNTN) that is based on deep learning advances for Arabic sentiment analysis.

3 BACKGROUND

3.1 The Iraqi Arabic Dialect

Iraq consists of a community structure in which different languages are spoken. The majority speaks Arabic, Kurdish, Turkmen and other languages. Iraqi dialect is often close to MSA and the Iraqis can pronounce the MSA sound of the vocal exits.

Our research focused on speakers of the Iraqi Arabic Dialect (IAD). IAD is the accumulation of several linguistic layers that have passed through the history of Iraq: Sumerian, Akkadian, Babylonian, Assyrian, Aramaic, and then Arabic. In addition to acquiring the vocabulary of Kurdish, Persian, Turkish and English, IAD uses many words that do not exist in MSA. These words are borrowed from other languages or adapted from the MSA. These words, although frequently used in everyday life, cannot be found in standard Arabic books. With the proliferation of online communities this dialect has taken on a broad scope as written texts in the Internet environment [18].

3.1.1 Phonology. IAD consists of three sub-dialects: **Moslawi, Baghdadi and Southern** [19]. The Moslawi dialect is spread in northern Iraq (Mosul city), also Tikrit city is close to this dialect. This dialect is close to MSA in terms of the pronunciation MSA /q/ phoneme (corresponding the letter ق /q/), which is pronounced as /g/ in Baghdadi and Southern dialects. **Baghdadi** dialect is used in the Iraqi capital Baghdad and some neighboring cities like Diyala, Samara, Anbar, Babylon and in some areas of Tikrit. It is characterized by simplicity, slow speech and clarity. **Southern** dialect is prevalent in the southern cities such as Missan, Basra, Dhi Qar, Qadisiyah, Karbala and Najaf. In **Southern** dialect, the pronunciation of **/k/** phoneme (corresponding the letter ك /k/**)** in MSA is converted to (ch) (corresponding the letter چ ch).

IAD has these additional consonants /g/, voiceless /ch/, and voiceless /p/ that are voiced more than MSA. Besides, there are two additional vowels: /e: / (ى), and /o: / (ۆ). They are longer than the classical sounds, though they have the same written forms [19].

3.1.2 Morphology. The words of the IAD do not end with vowel letters. It does not show the grammatical state of the word. Thus, unlike MSA, the words in IAD end with the consonants letters rather than vowels.

The past tense in IAD is influenced by the ancient languages of Mesopotamia and the Assyrian dialects of the original population of Iraq, where it begins with a 2- consonant cluster [19].

The prefix used to indicate the future tense in the Iraqi dialect is "راح / rāḥ/", which means "went". This "راح / rāḥ/" inserted before the present tense verb to indicate the distant future (will), while in MSA used "س" or "سوف" /swf /as prefix to the present verb to indicate the future tense.

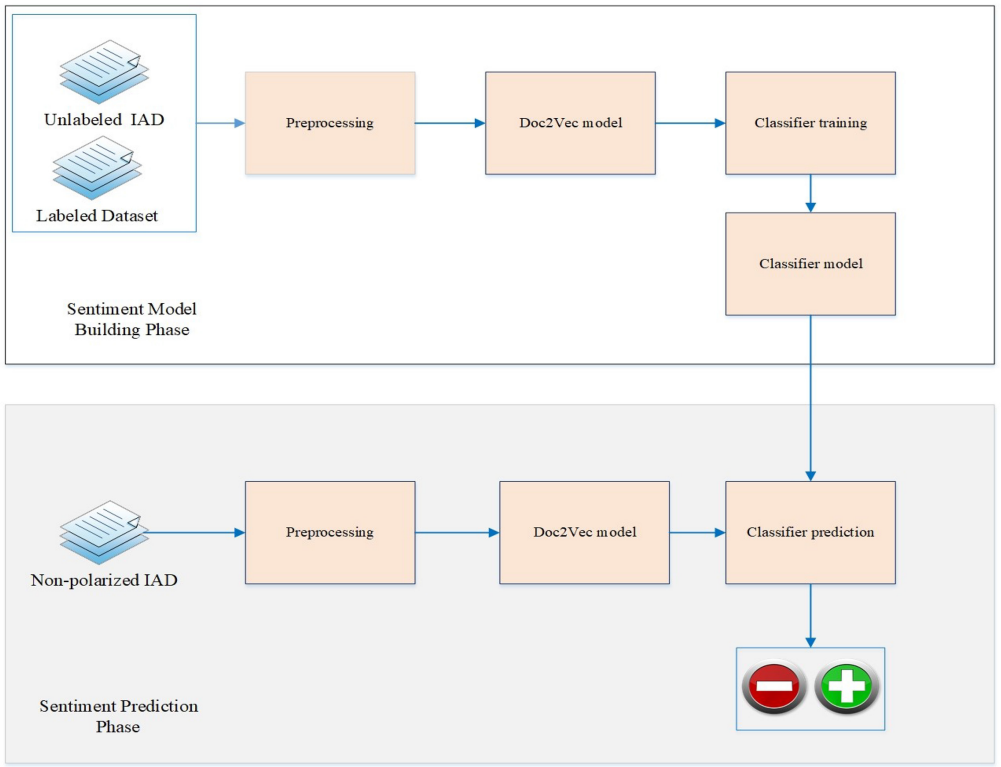

Fig. 1. IAD Sentiment analysis framework.

To denote the present tense, the prefix "دا /dā/" or "جاي /ğāy/" is added to the beginning of the verb, while the MSA does not have a form to refer to the present tense [19].

In IAD there is a single relative pronoun "اللي /ālly/" regardless of gender or number. The pronoun "اللي /ālly/" is derived from a MSA pronoun (الذي) that is used for singular masculine [19].

4 METHODOLOGY

With the limited available resources of labeled datasets of IAD, the proposed methodology assumes the use of labeled datasets in MSA and Arabic dialects (positive or negative) with un- labeled IAD. This methodology is related to the Iraqi dialect only and has not been tried on other dialects. Vector representation was extracted by training selected datasets. Four classification algorithms used to create sentiment prediction model, which can be used to predict new non-labeled documents. This methodology works with different types of vector representation and classification algorithms.

This methodology supports two main functions; building the model and prediction of sentiment.Building model stage, a set of classified documents is represents in vector space and train by classifier to build a prediction model. While at the prediction stage, the generated model was used to predict sentiment for the new given unlabeled document.

Figure 1 shows the layout of the proposed sentiments analysis approach for Arabic Iraqi dialect. The methodology of the proposed sentiment analysis framework assumes the existence of a labeled collection of documents, which belongs to the same domain and is typically labeled (positive or negative) based on the respective opinions ex-pressed.

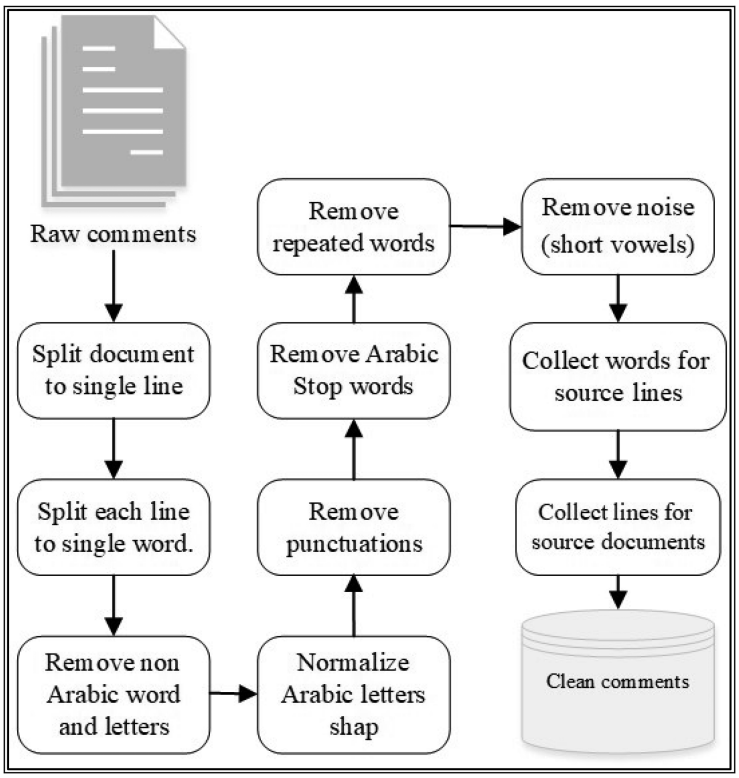

Fig. 2. Preprocessing steps.

4.1 Preprocessing

Data coming from the Internet contain a lot of noise and they are not structured; thus, we cleaned these data by following the steps in Figure 2.

Natural Language Processing (NLP) is one of the important techniques in the field of text mining. It covers computer understanding and manipulation of human language. As mentioned in Figure 2. many processes in the NLP will applied in preprocessing. First step, tokenized text file into smaller pieces. In our work this step performed in two sub-steps, split document to single line and split each line to single word. Further processing can be achieved after a text has been suitably tokenized. Our work aims Arabic text; thus, non-Arabic words will be deleted in second step. Arabic letters can be written in many forms. Normalization of Arabic letters will be performed after the removal of the punctuations and noise (short vowels) to allow processing to proceed uniformly. Stop words are not effected on sentiment decision; thus, they will be deleted in next step. Our approach is based on binary sentiment analysis. This means, the strength of the sentiment in the sentence is not required. The repeated words will not be effect on binary sentiment analysis, all of them will be deleted. The last two steps contain the aggregation of words to source line and each line to source document.

4.2 Word Embedding Models

Word embedding models are effective alternatives to build vector representation for texts [20, 21]. The word embedding represents the semantic information of words in the word-embedding space by learning vector representations of words [22]. Word2Vec model was developed by Google in

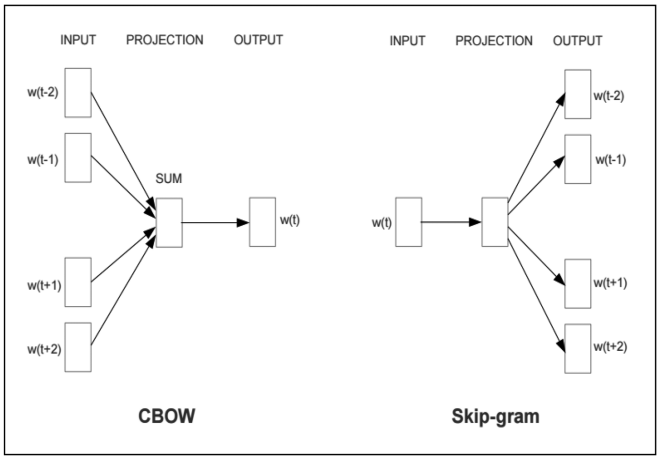

Fig. 3. CBOW and Skip-Gram models [20].

2013. In this model, the representation of words with the same semantic meaning is mapped nearby each to other in space. This model contains two architectures; Continuous Bag-Of-Words (CBOW) and Skip-Gram.

The input of the CBOW model could be $wt-2$, $wt-1$, $wt+1$, $wt+2$. The model receives a window of n words around the target word wt at each time step. Therefore, the CBOW model use $wt-2 \ldots wt+2$ words to predict wt. CBOW is trained to predict the word based on the given context. Skip-Gram is the opposite of this. The input of the Skip-Gram model is wt, and the output might be $wt-1$,$wt-2$,$wt+1$, $wt+2$. Skip-Gram is trained to predict the context based on the given word. Figure 3 shows the difference between CBOW and Skip-Gram.

We used Doc2Vec [23] which is based on Word2Vec implementation of word embedding. After applying our preprocessing, we started modeling our data in vector space by formatting each sentence as [['w_1', 'w_2', 'w_3',..., 'w_n'], ['label1']] [24]. Then, building the vocabulary table by embedding all words and filtering out the unique words, also counts words in a simple way. Now, we represent every word as vector in the space. This will be done by applying Doc2Vec. As a result, we have three clusters in the space; negative words, positive words and unlabeled words. Words that have the same contexts presented as closer as to one another in the space [25]. Therefore, unlabeled words will be close to positive or negative words. Figure 4 shows the word embedding model.

Doc2Vec contains two architectures models: Distributed Bag of Words (DBOW) and Paragraph and Distributed Memory Model of Paragraph Vectors (DM-PV). DBOW is similar to skip-gram. In this work we used DM-PV to learn the embedding; because, it is faster to train (than DBOW), simpler and suitable for larger datasets [23, 26, 27]. DM-PV model consists of three layers: the input layer which corresponds to the context, the hidden layer which processes each word from the input layer by projection into the weight matrix and output layer. Furthermore, to correct the representation of the word, it compares the output with the word itself based on the back propagation of the error gradient. DM-PV neural network aims to maximize the following formula [28]:

$$\frac{1}{V}\sum_{t=1}^{V}\log p(m_t|m_{t-\frac{c}{2}} \ldots m_{t+\frac{c}{2}}).$$

Where V represents vocabulary size, c represents window size of each word.

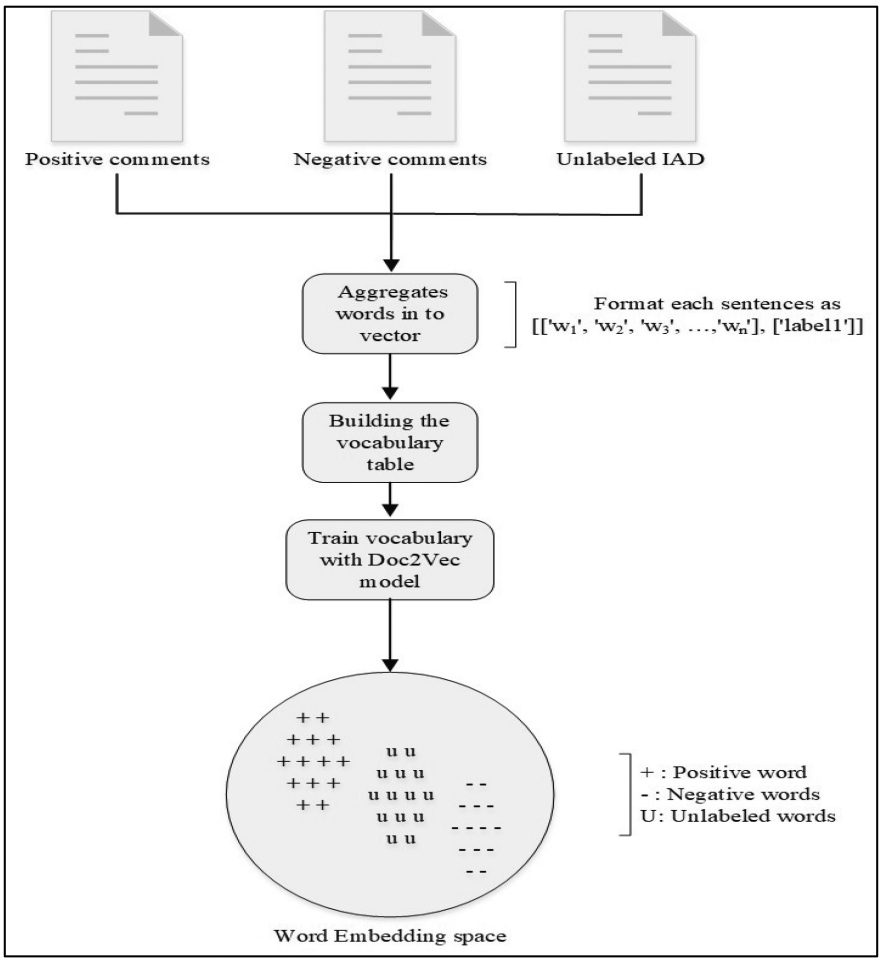

Fig. 4. Word Embedding Model.

4.3 Classifiers

For the training and test tasks, we applied classifiers that are widely used in the research area of the sentiment analysis [3, 12, 29, 30, 31]. LR, DT, SVM and NB are used with default parameters settings. For LR, we used the binary logistic regression. For DT we used default algorithm model. For SVC we used Linear SVC model. For NB we used Bernoulli NB.

5 EXPERIMENTS

In this section, the proposed sentiment detection approach evaluated by a series of experiments on selected datasets. We showed details on the datasets that used, experimentation approach, the settings of experiments, and finally the performance of methodology in terms of precision (denoted as P), recall (denoted as R) and F1 score (denoted as F1).

5.1 Experimental Setup

The evaluation of the proposed approach requires a variety of documents in different dialects, different lengths and domains. Therefore, we used a set of documents available in Arabic in both

Table 1. Corpus Collections and Sources

Dataset	Word count	Unique Word count	Positive Reviews	Negative Reviews
ATT [13]	61075	15041	1939	80
HTL [13]	894848	80454	10049	2470
RES [13]	295790	43022	7568	2513
MOV [13]	154978	36016	399	135
PROD [13]	35241	9975	2759	786
Twitter [31]	18404	8045	1000	1000
OCA [32]	130495	33964	250	250
ASTD [33]	21060	10622	799	1684
LABR [12]	694440	96489	42832	8224
MPQA [34]	163613	24600	2718	4911

MSA and dialects. In this paper, we will test the available corpora to classify the IAD, as well as the parameters (windows size, dimension and negative samples) of Doc2Vec model which may lead to a change in the classification performance. We addressed three questions about IAD sentiments detection:

- **RQ1**: By using Doc2Vec model as features representation, which parameters of Doc2Vec model will give a high performance?
- **RQ2**: By using Doc2Vec model as features representation, which classifier will give a high performance?
- **RQ3**: By using unlabeled dataset that content IAD for Doc2Vec training step, does the learned word embedding model improve the classification performance?

We addressed the answers for RQ1 in Section 6.1.1, RQ2 in Section 6.1.2 and RQ3 in Section 6.1.3

5.2 Dataset

For our experiments, we built our corpus from six of publicly available datasets that were explained in Table 1. The first dataset is Large Arabic Multi-domain Resources for Sentiment Analysis from[13], where the datasets cover five domains (positive and negative) as follows: Dataset of Attraction (ATT), Hotel (HTL), Restaurant (RES), Movie (MOV) and Product (PROD) Reviews. The second dataset is Twitter Data set for Arabic Sentiment Analysis from [31], which consists 2000 labelled tweets (positive tweets and negative) on multiple topics such as: politics and arts. These tweets were written in both (MSA) and the Jordanian dialect. The third dataset is OCA from [32], which consists 500 comments of movies (positive and negative) collected from several blogs. The forth dataset is ASTD: Arabic sentiment tweets dataset from [33]. The fifth dataset is LABR: Large Scale Arabic Book Reviews from [12], which consists of over 51,000 book reviews (positive and negative). The sixth dataset is Multi-Perspective Question Answering (MPQA) for Arabic from [34].

The dataset comes in CSV format with their associated binary sentiment polarity labels, positive and negative. The Twitter and OCA datasets come in two directories. One contains negative comments and the other contains positive comments. The other datasets come in one text file for each one. The text files have multi lines and the structure of lines could be like (comments, sentiment). The sentiment labels are presented as “1” for positive and “0” for negative. Some datasets are presented in labels as “1” for positive and “-1” for negative. We used simple python program

Table 2. An Examples of Training Datasets

Dataset	Positive comments	Negative comments
ATT [13]	المجمع راقي وجميع سبل الترفية والماركات موجودة فية. The mall is sophisticated and there is many entertainment and brands.	الجو حار كثيروالاقسام سيئة جداً والمبنى صغيرولا انصح بالذهاب إليه. It is so hot, the sections are very bad, the building is small, and I do not recommend going there.
HTL [13]	الموقع جميل جداً فندق بعيد عن الزحمه وقريب على كل الاماكن السياحيه وسعر معقول. The location is very nice and the hotel is far from crowded and close to tourist destinations and reasonable price.	في الواقع، إنه لا ينبغي أن يكون حتى فندق. إنه متهالك. In fact, it should not even be a hotel. It is worn out.
RES [13]	المنيو وجدته أسعاره في متناول الجميع والطعام جيد جدا. Affordable prices and a very good food.	سيء جداً لا انصح به اطلاقاً، أكل سيء الطعم، طريقه التقديم أسوأ. Very bad, I do not recommend it at all, tasteless food, The service worse.
MOV [13]	موسيقى هانز زيمر ممتازة و مناسبة لأجواء الفيلم. Hans Zimmer's music is good and suitable for the film.	اخراج سئ وتمثيل مبالغ فيه من أسوأ الافلام الي شفتهم في حياتي. Bad directing, bad acting, one of the worst films I have ever seen.
PROD [13]	تم توصيل السلعة الى المنزل بحالة ممتازه في وقت قصير The item was delivered to the home in excellent condition in a short time.	المنتج غير جيدة ولا انصح به The product is not good and I would not recommend it.
Twitter [31]	كلمات رائعه و حلوه و لا احلى ان يكونو في النهايه رائعون. Wonderful words and sweet and it is a great to be in the end a wonderful.	برنامج فاشل جدا. A very unsuccessful TV program.
OCA [32]	الفلم رائع من حيث احداثه واساليب تصويره. The film is wonderful in terms of events and methods of photography.	اسوء افلام الموسم. The worst movies of the season.
ASTD [33]	فكرة المقالة حلوة وأسلوبك في توصيل الفكرة حلو. The idea of the article and is good and	الراتب مايكفي الحاجة. Salary is not enough.
LABR [12]	رواية رائعة بما تحمله الكلمة من معنى . مراد تحسن اسلوبه كثيرا بعد فيرتيجو. A wonderful novel, Murat improved his style a lot after Vertigo.	كتاب سي جدا .الاسلوب غير ممتع ,نهاية مفتوحة والكتاب بوجهة عام كئيب. A very bad book. The writing style is not interesting and gloomy.
MPQA [34]	الحكومة الاندونيسية قد اتخذ خطوات جادة لضمان السلامة الشخصية للمستثمرين. The Indonesian government has taken serious steps to ensure the personal safety of investors.	ان الوضع هو الأسوأ في زيمبابوي في الجنوب. The situation is the worst in south of Zimbabwe.

to split lines into sentiment category. Consequently, each dataset becomes two files (positive and negative). All positive files are combined together, also negatives files. This datasets are referred as raw. Further processing is applied as suggested in section 4.1. Table 2 presents an examples of comments.

Table 3. Negative and Positive Comments in Each Domains for Test Data

Facebook pages	Positive comments	Negative comments
Home appliances company	306	170
Airways company	133	155
News page	71	170
Restaurant	216	16
Sport	145	147
Communications company	129	342
Sum	1000	1000

To cover the problem of the Iraqi dialect, we collected dataset that content Iraqi Arabic Dialect from Facebook. We chose Iraqi pages in different domains: home appliances company, Airways Company, News, Restaurant, Sport and Communications Company,

We used Facepager[1] to extract these data. By three native experts, we tagged 2000 comments manually for their sentiment (1000 positive and 1000 negative). These 2000 comments will be used as test data to evaluate the prediction model.

Unlabeled dataset that consists IAD (more than 250000 comments and 19000 unique words) also used for Doc2Vec training step. We will show the effects of unlabeled dataset on results in Section 6.1.3. Table 3 shows the content of the test dataset.

Formatting sentences and words is necessary before generating word embedding. Although the texts were carefully prepared in previous works, we noticed some irregularities in the texts such as the punctuation marks, repetition of letters in the word and some non-Arabic characters. Thus, we performed further preprocessing on the fully text datasets. Python's Natural Language Toolkit (NLTK)[2] was used as a tool for preprocessing because NLTK provides perfect assist dealing with texts in encoding issues [35]. The preprocessing includes the following steps:

- Delete non-Arabic characters.
- Delete punctuations.
- Stop word removal.
- Removes short vowels and other symbols (harakat)[3]
- Normalization: unifies the orthography of *alif*s, *hamza*s, and *yas*/*alif maqsura*s[4].

5.3 Building Word Embedding Models

We conducted our experiments with several settings for the parameters. With each setting of the parameters, we applied the generator embedding model to the sentiment classifiers to test the performance. Experimental settings have been investigated as suggested in the literature, including window sizes of 1, 2 and 3 [36, 37], embedding dimensions of 50, 100 and 200 [38, 39] and negative samples of 7, 10, 20 and 30 [20, 39]. We conducted our experience on both the implementation of Doc2Vec [23] modified by Linan Qiu[5] and Python's Gensim[6]. Finally, we made some modifications to suit the model with Arabic encoding and the data set for training and testing.

[1] https://github.com/strohne/Facepager.
[2] http://www.nltk.org/.
[3] https://maximromanov.github.io/2013/01-02.html.
[4] https://maximromanov.github.io/2013/01-02.html.
[5] https://github.com/linanqiu/word2vec-sentiments.
[6] https://radimrehurek.com/gensim/models/doc2vec.html.

Table 4. Results of the Classifiers Score Using Different Parameters (W = 1, 2, 3, D = 50)

	LR			DT			SVM			NB		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
W = 1	0.79	0.74	0.73	0.62	0.62	0.61	**0.81**	**0.76**	**0.75**	0.66	0.62	0.6
W = 2	0.79	0.75	0.74	0.62	0.61	0.6	0.79	0.72	0.71	0.68	0.64	0.61
W = 3	0.79	0.74	0.73	0.63	0.62	0.62	0.79	0.71	0.68	0.65	0.59	0.54

Table 5. Results of the Classifiers Score Using Different Parameters (W = 1, 2, 3, D = 100)

	LR			DT			SVM			NB		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
W = 1	0.79	0.74	0.73	**0.64**	**0.63**	**0.62**	0.8	0.73	0.72	0.67	0.64	0.62
W = 2	0.77	0.72	0.71	0.6	0.59	0.59	0.78	0.69	0.67	0.66	0.61	0.58
W = 3	0.78	0.73	0.72	0.59	0.58	0.58	0.79	0.69	0.66	0.64	0.62	0.6

Table 6. Results of the Classifiers Score Using Different Parameters (W = 1, 2, 3, D = 200)

	LR			DT			SVM			NB		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
W = 1	**0.8**	**0.77**	**0.76**	0.61	0.61	0.6	0.8	0.73	0.71	**0.71**	**0.66**	**0.65**
W = 2	0.79	0.75	0.74	0.58	0.58	0.57	0.79	0.69	0.66	0.67	0.64	0.63
W = 3	0.79	0.75	0.74	0.6	0.6	0.59	0.79	0.68	0.65	0.67	0.65	0.64

5.4 Evaluation Protocol

We considered that the standard sentiment classification is a binary classification. We instantiated our work with four typical classifiers: LR, DT, SVM and NB. All of the classifiers run under the same training conditions. In terms of N-fold cross validation, our datasets are already split into training and test datasets, for that no cross validation is use [26]. The evaluated performance of the classifiers in terms of measures, P, R and F1.

6 RESULTS

In the following sections, we reviewed the results derived from applying the classifiers on the generated word embedding model. First, we experimented with different parameters of Doc2Vec model and provided an answer to RQ1. Second, we experimented with four classifiers and gave an answer to RQ2. Third, we selected the best results of word embedding (in terms of parameters and classifiers) and tested it without using an unlabeled IAD dataset to show the effect of this dataset on the word embedding model and answered RQ3.

6.1 Word Embedding Parameters

In this section we investigated the effect of parameters (Window size, Dimensionality and Negative Samples) to deal with **RQ1**.

6.1.1 Effect of Window Size and Dimensionality. The measurements for the classifiers were arranged by window size of word embedding. For each row of window (1, 2 and3) and dimensionality (50,100, 200), Table 4, Table 5 and Table 6 show the results for window sizes. The best results were highlighted in bold.

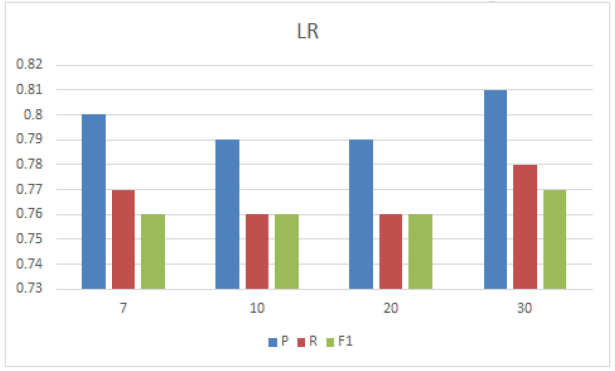

Fig. 5. LR classifier score using different Negative Sample sizes.

In Table 4, the results show there are no effects of changing windows size on P for LR classifier with little increase in R and F1 using W = 2. DT classifier show the high results for P, R and F1 with W = 3. SVM clearly show the best results in terms of P, R and F1 with W = 1. NB achieved the best result with W = 2.

Table 5, if we fix the dimensionality from 50 to 100, DT classifiers show improvement on P, R and F1 with W = 1. LR show little decrease in results in term of P, R and F1 with W = 1, 2. As well as a slightly change in result for both SVM and NB with W = 2.

The best results for LR and NB in terms of P, R and F1 was with W = 1 as show in Table 6. Comparing with Table 5, further we saw the results decreased slightly for DT with W = 1, 2.

In summary, for the IAD sentiments analysis task using word embedding with small context window size can give best results. However, the dimension size can affect the results based on a classifier that used.LR and NB get best results with large size of dimensionality.

Therefore, we answered RQ1, that for IAD sentiments analysis task, a small context window size was preferred which is different to that from other tasks.

6.1.2 Effect of Negative Samples. Negative samples size, is another important parameter that can affect the performance of the classification. When negative samples are used to train the model of word embedding, the number of negative samples are randomly determined for each data. Where the negative samples help model to differentiate between the correct words relationships from the noise [40]. indicated that big number of negative sample size is useful for small training set while for a large training set negative sample size can be as small as. In this part we illustrated the effect of negative sample sizes. For this, we trained the model using NS = (7, 10, 20, and 30) to cover a value from each range. Besides, we used a setting for dimensionality and window that attains good performance in previous section for experiments in this section.

From Figure 5, we observed that LR classifier show the best results with (W = 1, D = 200 and NS = 30) in terms of P, R and F1 (0.81, 0.78, 0.77).

From Figure 6, we observed that DT classifier show best results with (W = 2, D = 50 and NS = 10) in terms of P, R and F1 (0.66, 0.65, 0.65).

From Figure 7, we observed that SVM classifier show the best results with (W = 1, D = 50 and NS = 30) in terms of P, R and F1 (0.82, 0.79, 0.78).

From Figure 8, we observed that NB classifier show the best results with (W = 3, D = 100 and NS = 20) in terms of P, R and F1 (0.71, 0.66, 0.65).

In summary, by experimenting our dataset with different parameter values, we noted important effects of Negative Sample sizes on the performance of the classifiers. Where the best results were

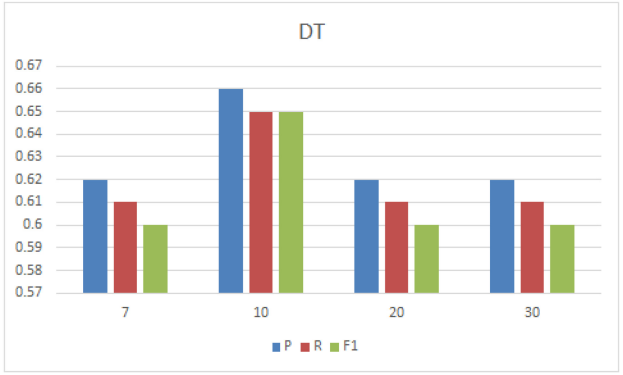

Fig. 6. DT classifier score using different Negative Sample sizes.

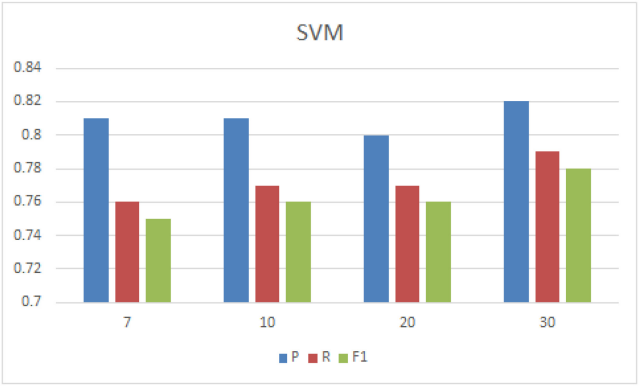

Fig. 7. SVM classifier score using different Negative Samples size.

with NS = 30 for LR, SVM and NB. For this, we can answer RQ2 that SVM get the best performance with value of parameters (W = 1, D = 50, and NS = 30).

6.1.3 Effect of Unlabeled IAD. We examined two cases for deriving the word embedding- based vectors. The first one involved training a Doc2Vec model on the given training data without unlabeled IAD. Second one using unlabeled IAD in space vector to derive the model. After generating the models, we derived vectors and applied the model building and sentiment prediction phases using four classifiers. In the two cases, a Doc2Vec model was generated using the parameter values that get best results for classifiers from Sections 6.1.1 and 6.1.2.

Table 7 clearly shows there is a difference in the performances of the classifiers. By using IAD, the vector value in space will be effective. This answers our RQ3 about using unlabeled IAD with dataset to generate Doc2Vec model does affect the performance of the LR, DT, SVM and NB classifiers.

7 DISCUSSION

Our experiments focused primarily on study the IAD sentiment analysis based on word embedding as feature representation in vector space. We made baseline for IAD sentiment analysis and compared them with other studies such as [36, 37] which concluded their studies with ‘a smaller

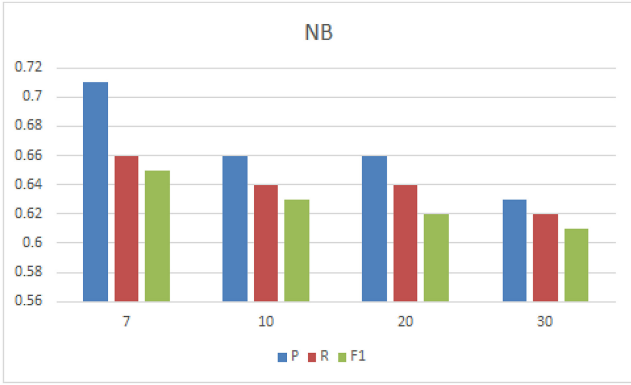

Fig. 8. NB classifier score using different Negative Sample sizes.

Table 7. Classifiers Performants with/without IAD.

Classifiers	Model	P	R	F1	TP*	TN**
LR	Without IAD	0.71	0.68	0.66	710	665
	With IAD	**0.81**	**0.78**	**0.77**	**810**	**771**
DT	Without IAD	0.60	0.60	0.60	600	600
	With IAD	**0.66**	**0.65**	**0.65**	**660**	**640**
SVM	Without IAD	0.77	0.74	0.74	770	729
	With IAD	**0.82**	**0.79**	**0.78**	**820**	**782**
NB	Without IAD	0.68	0.63	0.62	680	600
	With IAD	**0.71**	**0.66**	**0.65**	**710**	**634**

*True Positive (total positive comment = 1000).
**True Negative (total negative comment = 1000).

context window size gives a better performance''. We found our results similar to their conclusions and noticed that the best performance for our classifiers with window size is equal to 1.

Apart from this, the experimental in [3, 9] used a window size that is equal to 10 gives a better performance. Moreover, we studied the effect of dimensionality on sentiment analysis tasks and we noticed that there is a difference in performance between classifiers. SVM achieved the best performance with a dimensionality which is equal to 50, DT with a dimensionality which is equal to 100 and LR and NB with a dimensionality which is equal to 200. Therefore, we think the best setup of parameters and dimensionality may differ from task to another based on background corpora, balance of polarity and classifiers that are used. One of the important parameter that is not mentioned in [3, 35] is Negative Samples. In our experiments a larger Negative Samples gives slightly better performances for LR and SVM classifiers and vice versa for DT and NB classifiers. The important observation, by using unlabeled IAD, we got an improvement in the performances for all classifiers. That will contribute future studies on dialects that do not have own labeled corpus. Unlabeled words will be distributed to the nearest labeled word based on semantic in the vector space.

8 CONCLUSIONS

In this paper, we introduced word embedding as features extracting model for IAD sentiments analysis. For this purpose, we built a large Arabic corpus to generate word representations. We also

exploited word representations in space method to compute continuous vector representations of words using DBOW and studied the impact of word embedding parameters such as the context window, dimensionality and negative samples. Then, we applied four popular machine learning methods (LR, DT, SVM and NB). Our findings show that Negative Samples can indeed affect the classification performance. Besides, the best performance results of the classifiers was obtained using SVM with a high level of P (82 %) and R (79%) and F1 score (78%) when the classifier was applied by using the parameter values (W = 1, D = 50 and NS = 30).

REFERENCES

[1] Thabit Sabbah, Ali Selamat, Md Hafiz Selamat, Fawaz S. Al-Anzi, Enrique Herrera Viedma, Ondrej Krejcar, and Hamido Fujita. 2017. Modified frequency-based term weighting schemes for text classification. *Applied Soft Computing* 58 (September 2017), 193–206. DOI: https://doi.org/10.1016/j.asoc.2017.04.069

[2] Hassan Saif, Yulan He, Miriam Fernandez, and Harith Alani. 2016. Contextual semantics for sentiment analysis of twitter. *Information Processing & Management* 52, 1 (January 2016), 5–19. DOI: https://doi.org/10.1016/j.ipm.2015.01.005

[3] A. Aziz Altowayan and Lixin Tao. 2016. Word embeddings for arabic sentiment analysis. In *Proceedings of the IEEE International Conference on Big Data*. IEEE, Los Alamitos, CA, USA, 3820–3825. DOI: https://doi.org/10.1109/BigData.2016.7841054

[4] Bing Liu. 2012. *Sentiment Analysis and Opinion mining*, Morgan & Claypool Publishers. California, USA. DOI: https://doi.org/10.2200/S00416ED1V01Y201204HLT016

[5] Aymen Abu-Errub, Ashraf Odeh, Qusai Shambour, and Osama Al-Haj Hassan. 2014. Arabic roots extraction using morphological analysis. *International Journal of Computer Science Issues (IJCSI)* 11, 2 (March 2014), 128–134.

[6] Alaa M. El-Halees. 2017. Arabic opinion mining using distributed representations of documents. In *Proceedings of the Palestinian International Conference on Information and Communication Technology*. IEEE, Washington, DC, USA, 28–33. DOI: https://doi.org/10.1109/PICICT.2017.15

[7] RM Duwairi, Nizar A. Ahmed, and Saleh Y. Al-Rifai. 2015. Detecting sentiment embedded in arabic social media–a lexicon-based approach. *Journal of Intelligent & Fuzzy Systems* 29, 1 (2015), 107–117. DOI: https://doi.org/10.3233/IFS-151574

[8] Abdullateef M. Rabab'ah, Mahmoud Al-Ayyoub, Yaser Jararweh, and Mohammed N. Al-Kabi. 2016. Evaluating sentistrength for arabic sentiment analysis. In *Proceedings of the 7th International Conference on Computer Science and Information Technology (CSIT)*. IEEE, Washington, DC, USA, 1–6. DOI: https://doi.org/10.1109/CSIT.2016.7549458

[9] Sadam Al-Azani and El-Sayed M. El-Alfy. 2017. Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short arabic text. In *Proceedings of the 8th International Conference on Ambient Systems, Networks and Technologies, ANT 2017. Procedia Computer Science*, 359–366. DOI: https://doi.org/10.1016/j.procs.2017.05.365

[10] Anwar Alnawas and Nursal Arıcı. 2018. The corpus based approach to sentiment analysis in modern standard arabic and arabic dialects: A literature review. *Journal of Polytechnic* 21, 2 (June 2018), 461–470. DOI: https://doi.org/10.2339/politeknik.403975

[11] Abdelghani Dahou, Shengwu Xiong, Junwei Zhou, Mohamed Houcine Haddoud, and Pengfei Duan. 2016. Word embeddings and convolutional neural network for arabic sentiment classification. In *Proceedings of the 26th International Conference on Computational Linguistics*. Association for Computational Linguistics, Stroudsburg PA USA, 2418–2427.

[12] Mohamed Aly and Amir Atiya. 2013. Labr: Large scale arabic book reviews. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Stroudsburg, PA, USA, 494–498.

[13] Hady ElSahar and Samhaa R. El-Beltagy. 2015. Building large arabic multi-domain resources for sentiment analysis. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, Cham, Switzerland, 23–34. DOI: https://doi.org/10.1007/978-3-319-18117-2_2

[14] Eshrag Refaee and Verena Rieser. 2014. An arabic twitter corpus for subjectivity and sentiment analysis. In *Proceedings of the 9th International Language Resources and Evaluation Conference*. European Language Resources Association, France, 2268–2273.

[15] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*, Association for Computational Linguistics, Stroudsburg PA, USA, 746–751.

[16] Ahmad Al-Sallab, Ramy Baly, Hazem Hajj, Khaled Bashir Shaban, Wassim El-Hajj, and Gilbert Badaro.2017.Aroma: A recursive deep learning model for opinion mining in arabic as a low resource language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 16, 4, Article 25 (July 2017), 20 pages. DOI: https://doi.org/10.1145/3086575

[17] Ramy Baly, Hazem Hajj, Nizar Habash, Khaled Bashir Shaban, and Wassim El-Hajj. 2017. A sentiment treebank and morphologically enriched recursive deep models for effective sentiment analysis in arabic. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 16, 4, Article 23 (July 2017), 21 pages. DOI: https://doi.org/10.1145/3086576

[18] Fadi Biadsy, Julia Hirschberg, and Nizar Habash. 2009. Spoken arabic dialect identification using phonotactic modeling. In *Proceedings of the EACL 2009 workshop on computational approaches to semitic languages*. Association for Computational Linguistics, Stroudsburg, PA, USA, 53–61.

[19] Matti Phillips Khoshaba Al-Bazi. 2005. *Iraqi Dialect Versus Standard Arabic*, Matti Phillips Khoshaba (Al- Bazi). United States.

[20] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv:1301.3781. Retrieved from https://arxiv.org/abs/1301.3781.

[21] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Stroudsburg PA, USA, 1532–1543. DOI: https://doi.org/10.3115/v1/D14-1162

[22] Ahmet Hayran and Mustafa Sert. 2017. Sentiment analysis on microblog data based on word embedding and fusion techniques. In *Proceedings of the 25th Signal Processing and Communications Applications Conference (SIU)*. IEEE, Washington, DC, USA, 1–4. DOI: https://doi.org/10.1109/SIU.2017.7960519

[23] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*. JMLR.org, 1188–1196.

[24] Antoine J-P Tixier, Michalis Vazirgiannis, and Matthew R. Hallowell. 2016. Word embeddings for the construction domain. ArXiv:1610.09333. Retrieved from https://arXiv:1610.09333.

[25] Aitor García-Pablos, Montse Cuadros, and German Rigau. 2018. W2vlda: Almost unsupervised system for aspect based sentiment analysis. *Expert Systems with Applications* 91 (January 2018), 127–137. DOI: https://doi.org/10.1016/j.eswa.2017.08.049

[26] Maria Giatsoglou, Manolis G. Vozalis, Konstantinos Diamantaras, Athena Vakali, George Sarigiannidis, and Konstantinos Ch. Chatzisavvas. 2017. Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications* 69 (March 2017), 214–224. DOI: https://doi.org/10.1016/j.eswa.2016.10.043

[27] Sungwoon Choi, Jangho Lee, Min-Gyu Kang, Hyeyoung Min, Yoon-Seok Chang, and Sungroh Yoon. 2017. Large-scale machine learning of media outlets for understanding public reactions to nation-wide viral infection outbreaks. *Methods* 129, 1 (October 2017), 50–59. DOI: https://doi.org/10.1016/j.ymeth.2017.07.027

[28] Marwa Naili, Anja Habacha Chaibi, and Henda Hajjami Ben Ghezala. 2017. Comparative study of word embedding methods in topic segmentation. *Procedia Computer Science* 112 (September 2017), 340–349. DOI: https://doi.org/10.1016/j.procs.2017.08.009

[29] Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yaser Jararweh, and Omar Qawasmeh. 2018. Enhancing aspect-based sentiment analysis of arabic hotels' reviews using morphological, syntactic and semantic features. *Information Processing & Management* (January 2018). DOI: https://doi.org/10.1016/j.ipm.2018.01.006

[30] Hunaida Awwad and Adil Alpkocak. 2017. Using hybrid-stemming approach to enhance lexicon-based sentiment analysis in arabic. In *Proceedings of the International Conference on New Trends in Computing Sciences (ICTCS)*. IEEE, Los Alamitos, CA, USA, 229–235. DOI: https://doi.org/10.1109/ICTCS.2017.26

[31] Nawaf A. Abdulla, Nizar A. Ahmed, Mohammed A. Shehab, and Mahmoud Al-Ayyoub. 2013. Arabic sentiment analysis: Lexicon-based and corpus-based. In *Proceedings of the IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*. IEEE, Los Alamitos, CA, USA, 1–6. DOI: https://doi.org/10.1109/AEECT.2013.6716448

[32] Mohammed Rushdi-Saleh, M Teresa Martín-Valdivia, L. Alfonso Ureña-López, and José M. Perea-Ortega. 2011. Oca: Opinion corpus for arabic. *J. Am. Soc. Inf. Sci. Technol.* 62, 10 (October 2011), 2045–2054. DOI: http://dx.doi.org/10.1002/asi.21598

[33] Mahmoud Nabil, Mohamed Aly, and Amir Atiya. 2015. Astd: Arabic sentiment tweets dataset. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg PA, USA, 2515–2519. DOI: https://doi.org/10.18653/v1/D15-1299

[34] Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2010. Multilingual subjectivity: Are more languages better? In *Proceedings of the Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, 28–36.

[35] Matic Perovšek, Janez Kranjc, Tomaž Erjavec, Bojan Cestnik, and Nada Lavrač. 2016. Textflows: A visual programming platform for text mining and natural language processing. *Science of Computer Programming* 121 (June 2016), 128–152. DOI: https://doi.org/10.1016/j.scico.2016.01.001

[36] Fréderic Godin, Baptist Vandersmissen, Wesley De Neve, and Rik Van de Walle. 2015. Multimedia lab @ acl wnut ner shared task: Named entity recognition for twitter microposts using distributed word representations. In *Proceedings of the Workshop on Noisy User-generated Text*. Association for Computational Linguistics, Stroudsburg, PA, USA, 146–153. DOI: https://doi.org/10.18653/v1/W15-4322

[37] Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, 809–815. DOI: https://doi.org/10.3115/v1/P14-2131

[38] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Long Papers)*. Association for Computational Linguistics, Stroudsburg PA, USA, 1555–1565. DOI: https://doi.org/10.3115/v1/P14-1146

[39] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13)*. Curran Associates Inc., USA, 3111–3119.

[40] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. ArXiv:1309.4168. Retrieved from https://arxiv.org/abs/1309.4168.

Received June 2018; revised September 2018; accepted September 2018