The Use of Word2vec Model in Sentiment Analysis: A Survey

Conference Paper · December 2019

DOI: 10.1145/3388218.3388229

2 authors:


Samar Al-Saqqa
University of Jordan


Arafat Awajan
Princess Sumaya University for Technology

The Use of Word2vec Model in Sentiment Analysis: A Survey

Samar Al-Saqqa
Department of Computer Science
Princess Sumaya University for Technology
Amman, Jordan
King Abdullah II School for Information Technology
Department of Information Technology
The University of Jordan
Amman, Jordan,
s.alsaqqa@ju.edu.jo

Arafat Awajan
Department of Computer Science
Princess Sumaya University for Technology
Amman, Jordan
awajan@psut.edu.jo

ABSTRACT
Sentiment analysis is an area that gains wide interest from research because of its importance and advantages in various fields. Different approaches and techniques are used to classify the sentiment of texts. Word embedding is one of the effective methods that represent aspects of word meaning and help to improve sentiment classification accuracy. Word2vec is well-known and widely used in learning word embedding that includes two models: Skip-Gram (SG) model and Continuous Bag-of-Words model (CBOW). Some of the studies use one of these models and other studies use both of them. In this survey, we highlight the latest studies on using the Word2vec model for sentiment analysis and its role in improving sentiment classification accuracy.

CCS Concepts
• **Computing methodologies→Machine Learning**

Keyword
Sentiment Analysis, Word2vec, word embedding, deep learning, skip-gram, CBOW.

1. INTRODUCTION
Web content has become an integral part of our lives. It is rich with user-generated content forms, such as reviews, recommendations, blog posts, tweets, and comments. Daily user interactions with websites, social applications, and e-commerce sites increase the need to understand these vast sources of information by proposing new techniques and methods for how to handle and use this content for different purposes. An analysis of the sentiments of this content has become very useful in various areas such as business, education, and government. The process of detecting and determining whether a given text is positive, negative, or neutral is called Sentiment Analysis (SA). Sentiment analysis is a subfield of Natural Language Processing (NLP) which is a subfield of artificial intelligence that deals with human-computer interactions and explores how the computer can be used to understand a natural text and manipulate it for useful tasks. [1], [2].

Sentiment analysis has three main levels: document level, sentence level, and aspect level. The main task at the document level is to determine whether the entire opinion document expresses a negative or positive polarity. The sentence level is concerned with the sentiment of each sentence, and the aspect level is concerned with extracting sentiments from different aspects of objects [3], [30].

There is a lot of research that proposed numerous techniques to analyze and classify the sentiment of the text, these techniques can be categorized into three approaches: supervised, unsupervised, and hybrid. The supervised approach is a machine learning approach involving the use of classification algorithms such as Naïve Bayes and support vector machines, maximum entropy, etc. The unsupervised approach includes various methods that are based on sentiment lexicons. The hybrid approach combines both machine learning and lexicon-based approaches. [4]

Researchers have recently applied new techniques in sentiment analysis. Deep learning is one of the most common and powerful machine learning techniques that has been widely applied to sentiment analysis and shows great potentials and impacts on the performance of sentiment analysis [5], [6], [7], [31].

Deep learning is a machine learning method that uses artificial neural networks in learning tasks using networks of many hidden layers [5]. In NLP, Deep learning models need word embedding which is a type of word representation, the words are transformed into vectors. There are different models used for word embedding. One of commonly used word embedding models is Word2vec. [8], [9], [10].

In the literature there are various surveys on sentiment analysis and its techniques and approaches, however, to the best of our knowledge, there is no survey specifically addresses the use of the word2vec model in sentiment analysis. So in this survey, we focus on exploring the researches that were done using the word2vec model. This paper is organized as follows. We introduce a brief background of word embedding in section II. Word2vec model and its algorithms are introduced in section III.

AIRC'19, December 14–16, 2019, Cairo, Egypt

ACM ISBN 978-1-4503-7671-6/19/12…$15.00
DOI:https://doi.org/10.1145/3388218.3388229

The use of the word2vec model for sentiment analysis is presented in section IV and section V provides the conclusion.

2. WORD EMBEDDING

There is a lot of progress being currently made in NLP using word embedding, it is a positive trend that can be used in a very broad range of practical NLP applications such as computing the similarities between words, using as features in text classification and different natural language tasks such as sentiment analysis. [11]. Word embedding is used in sentiment analysis and it has good achievements and results on analysis performance [12], this solves the drawbacks of the Bag of Words (BOW) model. A document in BOW is converted to a numeric feature vector with a fixed length, each element of the vector can be word occurrence (absence or presence), word frequency or TF-IDF score. Its dimension equals to the size of the vocabulary [5]. Although widely used to analyze sentiment, it has its drawbacks. The first drawback of BOW is the loss of word order which means that two documents can have the same representation as long as they share the same words. The second one that BOW ignores the semantics of words [13], [15], [16], [17].

Word embedding is a technique for feature learning, in which the words of vocabulary are converted to vectors of continuous real numbers with low dimensions [5]. It is a distributional vector representation of the words and it is also called semantic vector space since it captures both the semantic (meaning) and syntactic (structure) information of words from the context in which they are used. Word embeddings are mostly based on neural networks. At first, the random initialization for word vectors happens, then the context is predicted optimally after vector training in which the corresponding words tend to appear and semantically similar words have similar vectors. There are two common approaches through which word vectors are generated, both of which are very related to each other, the first approach is based on word count or word context co-occurrences. The second approach is based on predicting the word based on its context.

The well-known models that are used for producing word embeddings in NLP are Word2vec and Global Vectors for Words Representation (GloVe) [9], [17]. Word2vec is a technique that was introduced by Mikolov et al. [9], it uses a shallow neural network to learn word embeddings. The glove is an unsupervised learning algorithm for obtaining vector representation of words proposed by Pennington et al. [13] does very well at context preservation. These two techniques have been reported to be effective in different natural language processing applications [9], [17], [28], [29], however, as seen from literature word2vec is the most commonly used for word embedding for sentiment analysis application and this is because of the good performance of word2vec [28]. In this survey, we focus on using the word2vec model in the sentiment analysis.

3. WORD2VEC MODEL

Word2vec is well known and widely used in learning word embedding from raw text introduced by Mikolov et al. [9]. The idea of word2vec (word embeddings) originated from the concept of distributed representation of words, it uses a shallow neural network to learn word embeddings and predicts between every word and its context words so words occurring in similar contexts are related.

There are two algorithms inside word2vec for producing word vectors, skip-gram and Continuous Bag-of-Words.

3.1 Skip-Gram Model (SG)

The idea of SG, in each estimation step you are taking one word as a center word, and you are going to try and predict words in its context out to some window size, and the model is going to define a probability distribution that is the probability of a word appearing in the context given this center word. And we are going to choose vector representation of words so we can try and maximize that probability distribution (we just have one probability distribution of the context word).

3.2 Continuous Bag-of-Words Model (CBOW)

In this model, the target word (center word) is predicted based on the surrounding words. The context is represented by multiple words, and to predict the missing target word, the sequence of context words must be given, then try to predict the missing word in the context according to the window size. This can be applied to different sizes of windows (10,15,20,100, etc.).

According to [9], CBOW works several times faster than SG to train and slightly better accuracy for frequent words, while SG works well with a small amount of training data, and represents well even rare word or phrases. Figure 1 shows the architecture of CBOW and SD algorithms.

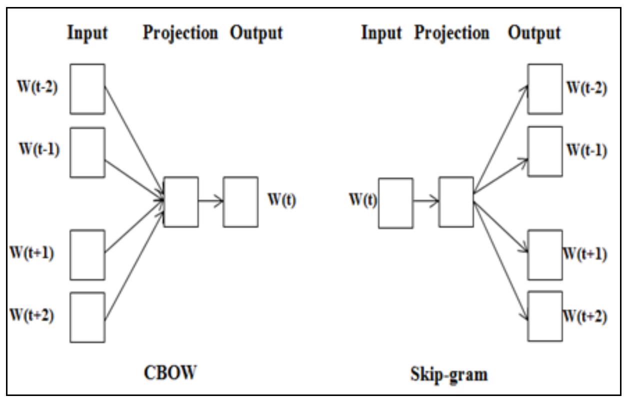

Figure 1. The architecture of CBOW and Skip-gram algorithms [9].

4. SENTIMENT ANALYSIS AND WORD2VEC

Lots of work have been done using word2vec, some research used CBOW or SG and others used both approaches. Most of the research has used different datasets and applied the Word2vec model in sentiment analysis on different languages as shown in Table 1 and Table 2.

Pouransari and Saman in [6], used different NLP methods to perform sentiment classification with binary and multiclass labels. For the binary classification, they applied a bag of words and word2vec skip-gram followed with various SVM, random forests and logistic classifiers. They used vector averaging, and clustering to aggregate word vectors into a single feature vector for each review. The results of binary classification using Word2vec with averaging for the random forest, SVM, and logistic regression were 84.0%, 85.8%, and 86.6% respectively. And the accuracy for word2vec and random forests with clustering was 83.5%.

Sentiment classification accuracy is improved by employing word embedding in [12], high-dimensional word vectors that learn Contextual information of words are produced by using word2vec. They used the Skip-Gram Model of word2vec to perform sentiment analysis of tweets relating to U.S. Military Base in Ghana. Random Forest classifier is used for training. To evaluate the performance, they used evaluation metrics accuracy, recall, precision, and F1-score metrics. The overall accuracy for the sentiment labels was 81% which indicates that the word vectors quality that was produced by the skip-gram model help to give good results in sentiment polarities prediction.

Table 1. The datasets that are used in literature

Ref no.	Dataset Name
[6]	Kaggle1 competition Dataset
[12],[27],[18]	Twitter Dataset
[14]	Sina Weibo
[15]	Chinese comments on clothing products.
[19]	ACL300,ACL100, Brown100
[20]	ASTD-ArTwitter,-QCRI, LABR, MPQA
[21]	U.S Airlines Tweets
[23]	Airline Tweets
[25]	LABR, ASTD, MPQA
[26]	Arabic Health Services dataset (AHS)

Table 2. Word2vec-Sentiment analysis is applied to different languages

Ref no.	Language
[6],[12],[18],[19],[21],[27]	English
[14],[15]	Chinese
[20],[23],[25],[26]	Arabic

Xue et al [14] proposed a new similarity distance model, Semantic Orientation Pointwise Similarity Distance (SO-SD) model, and they used Word2vec to create sentiment dictionary based on their proposed model, Emotional Dictionary was created to obtain the emotional tendencies of Weibo messages. The experiments showed good results using this approach. Based on word2vec and SVMperf, a new sentiment classification method is proposed in [15]. Using a data set of Chinese clothing products comments from Amazon, they applied the word2vec tool to cluster the similar features to show the word2vec capability to capture the semantic features of the Chinese language then comment texts were trained and classified again using word2vec again and SVMperf. Their findings concluded the superior performance of their method in sentiment classification.

Wang and Jorge [18] studied the relationship between sentiment classification, emoticons, and the contexts in which emoticons are used. To interpret emoticons in the context of tweets they used word2vec to define the representation of the words, including emoticons in the dataset. They used the k-means algorithm to cluster the words so the exact meaning of the emoticons and be understood through the words that appear in the same cluster. Liu in [19], Used the word2vec model to analyze citations sentiment. Word embedding for sentiment analysis of citations is evaluated. Sentence embedding is constructed based on word embedding; it is obtained by averaging the vectors of the words in one sentence which is trained using word2vec. The sentence embedding is evaluated using three datasets, the results revealed that word2vec is effective and promising on classifying positive and negative citations, however, hand-crafted features have a better performance.

Altowayan et al. [20], Proposed an alternative approach of hand-crafted features for Arabic sentiment analysis, this approach is neural word embedding to get distributed representations for words, the word representation is employed to embed features as dense vectors instead of the conventional sparse representations. They built a huge corpus from news articles, Quran text, and consumer reviews and then they used Continuous Bag-Of-Words (CBOW) model of word2vec to pre-trained Arabic word embedding and generate word vectors. The classification experiments were performed using six binary classifiers on three datasets. The results of experiments showed that the SVM classifier and logistic regression classifier performed better than others in terms of different evaluation metrics; macro-accuracy, recall, precision, and f-measure and the results revealed that their embedding approach gives better performance in comparison with other techniques.

Acosta et al. [21], employed word embedding generated by word2vec model to classify sentiment, they used both models of word2vec; skip-gram (SG) and continuous bag-of-words (CBOW) in their experiments. Four classifiers were trained with both models of word2vec, the classifiers are Gaussian Naive Bayes, Bernoulli Naive Bayes, Support Vector Machines and Logistic Regression. The performance of the support vector machine and logistic regression using the skip model outperforms the naïve bays classifiers.

Ashi et al. [23], compared two models of word embedding for Arabic Aspect-Based Sentiment Analysis, AraVec-Web which is built using the word2vec Skip-Gram technique [24] and fastText Arabic Wikipedia word embeddings. The compared results showed that the performance of fastText Arabic Wikipedia word embeddings is slightly better than AraVec-Web. FastText is also used in [25], it is used for both skip-gram and continuous bag-of-words models in training and building sentiment-specific embeddings. After training and testing five classifiers on the three selected datasets, the results revealed that CBOW and skip-gram models perform well on syntactic and semantic analogies. Yu et al. [32], proposed refinement model for word vectors which can be applied by word2vec pre-trained word vectors. The model is based on selecting similar nearest neighbors in semantic and ranking them in descending order from similar neighbors to dissimilar neighbors, these ranked set to guide the refinement in a manner that helps in improving the vector representation of the words and word embedding performance. Table 3 shows a summary of articles using the word2vec model in the sentiment analysis.

Table 3. Word2vec models used for sentiment analysis

Ref no.	Year	Model	Evaluation Metric	Approach
[6]	2014	Skip-gram	Accuracy	Random forest, support vector machines(SVM) , Logistic Regression.
[12]	2018	Skip-gram	Precision, recall and F1-score, accuracy.	Random Forest
[14]	2014	Skip-gram, Continuous Bag-Of-Words	Precision, Recall, and F-measure.	N/A
[15]	2014	Continuous Bag-Of-Words	Precision, Recall, and F-measure	Support Vector Machines (SVM^{perf})
[18]	2015	Skip-gram, Continuous Bag-Of-Words	Accuracy, Precision Recall F1-score	Naïve Bayes
[19]	2017	Classical Word2Vec algorithm(Mikolov)	macro-F and weighted-F	Support Vector Machines(SVM)
[20]	2016	Continuous Bag-Of-Words	macro-accuracy, Precision, Recall, and F-measure, and	Support Vector Machines(SVM), Gaussian Naive Bayes, SGD Classifier, Logistic Regression
[21]	2017	Skip-gram, Continuous Bag-Of-Words	Accuracy, Precision Recall F1-score	Support Vector Machines (SVM), Gaussian Naive Bayes, Bernoulli Naive Bayes, Logistic Regression
[23]	2018	AraVec-Web, (Word2Vec Skip-gram)	Precision, Recall, and F-measure, accuracy,	SVM linear classifier
[25]	2017	Skip-gram, Continuous Bag-Of-Words, FastText	F1-score	Gaussian Naive Bayes, Random Forest, Support Vector Machines(SVM), Stochastic Gradient Descent, Logistic Regression
[26]	2018	Skip-gram, Continuous Bag-Of-Words	Accuracy	Stochastic gradient descent, logistic regression, Bernoulli Naive Bayes, Multinomial Naive Bayes, Nu-Support Vector classification, linear support vector Ridge classifier
[27]	2016	Word2Vec algorithm(Mikolov)	Accuracy	Random Forest

5. CONCLUSION

Word2vec is one of the common models that is used in a word embedding. It is a prediction based algorithm that is used to represent a word as vectors with semantic relation. It is applied in many NLP tasks and has shown a great potential impact on the performance of sentiment analysis. In this survey, we presented a literature review of several studies that used the word2vec model for sentiment analysis. It can be seen from the literature that most studies were used the two methods of word2vec: CBOW and skip-gram, and compared the results from each method. Skip-gram is better for infrequent words than CBOW, however, CBOW is faster and works well with frequent words. Many of the studies in literature applied word2vec using tools such as word2vec tool and FastText.

6. REFERENCES

[1] Hirschberg, Julia, and Christopher D. Manning. Advances in natural language processing. *Science* 349, no. 6245 (2015): 261-266.

[2] Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." *Ain Shams Engineering Journal* 5, no. 4 (2014): 1093-1113.

[3] Feldman, Ronen. Techniques and applications for sentiment analysis. *Communications of the ACM* 56, no. 4 (2013): 82-89.

[4] Ravi, Kumar, and Vadlamani Ravi. A survey on opinion mining and sentiment analysis: tasks, approaches and applications."Knowledge-Based Systems 89 (2015): 14-46.

[5] Zhang, Lei, Shuai Wang, and Bing Liu. "Deep learning for sentiment analysis: A survey." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8, no. 4 (2018): e1253.

[6] Pouransari, Hadi, and Saman Ghili. Deep learning for sentiment analysis of movie reviews. Technical report, Stanford University, 2014.

[7] Ain, Qurat Tul, Mubashir Ali, Amna Riaz, Amna Noureen, Muhammad Kamran, Babar Hayat, and A. Rehman. Sentiment analysis using deep learning techniques: a review. *Int J Adv Comput Sci Appl* 8, no. 6 (2017): 424.

[8] Rong, Xin. Word2vec parameter learning explained. arXiv preprint arXiv:1411.2738 (2014).

[9] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space." arXiv preprint arXiv: 1301.3781 (2013).

[10] Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." In *Advances in neural information processing systems*, pp. 3111-3119. 2013.

[11] Lebret, Rémi Philippe. Word Embeddings for Natural Language Processing. No. THESIS_LIB. EPFL, 2016.

[12] Deho, B. Oscar, A. William Agangiba, L. Felix Aryeh, and A. Jeffery Ansah. "Sentiment Analysis with Word Embedding." In 2018 IEEE 7th International Conference on Adaptive Science & Technology (ICAST), pp. 1-4. *IEEE*, 2018.

[13] Barhoumi, Amira, Yannick Estève1 Chafik Aloulou, and Lamia Hadrich Belguith. "Document embeddings for Arabic

Sentiment Analysis." *Language Processing and Knowledge Management* (2017).

[14] Xue, Bai, Chen Fu, and Zhan Shaobin. "A study on sentiment computing and classification of sina weibo with word2vec." In Big Data (BigData Congress), 2014 IEEE International Congress on, pp. 358-363. *IEEE*, 2014.

[15] Zhang, Dongwen, Hua Xu, Zengcai Su, and Yunfeng Xu. "Chinese comments sentiment classification based on word2vec and SVMperf." *Expert Systems with Applications* 42, no. 4 (2015): 1857-1863.

[16] Lilleberg, Joseph, Yun Zhu, and Yanqing Zhang. "Support vector machines and word2vec for text classification with semantic features." In Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th International Conference on, pp. 136-140. *IEEE*, 2015

[17] Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532-1543. 2014.

[18] Wang, Hao, and Jorge A. Castanon. "Sentiment expression via emoticons on social media." arXiv preprint arXiv:1511.02556 (2015)

[19] Liu, Haixia. "Sentiment analysis of citations using word2vec." arXiv preprint arXiv: 1704.00177 (2017).

[20] Altowayan, A. Aziz, and Lixin Tao. "Word embeddings for Arabic sentiment analysis." In Big Data (Big Data), 2016 IEEE International Conference on, pp. 3820-3825. *IEEE*, 2016.

[21] Acosta, Joshua, Norissa Lamaute, Mingxiao Luo, Ezra Finkelstein, and Andreea Cotoranu. "Sentiment Analysis of Twitter Messages Using Word2Vec." (2017).

[22] Deho, B. Oscar, A. William Agangiba, L. Felix Aryeh, and A. Jeffery Ansah. "Sentiment Analysis with Word Embedding." In 2018 IEEE 7th International Conference on Adaptive Science & Technology (ICAST), pp. 1-4. *IEEE*, 2018.

[23] Ashi, Mohammed Matuq, Muazzam Ahmed Siddiqui, and Farrukh Nadeem. "Pre-trained Word Embeddings for Arabic Aspect-Based Sentiment Analysis of Airline Tweets." In *International Conference on Advanced Intelligent Systems and Informatics*, pp. 241-251. Springer, Cham, 2018.

[24] Soliman, Abu Bakr, Kareem Eissa, and Samhaa R. El-Beltagy. "Aravec: A set of arabic word embedding models for use in arabic nlp." *Procedia Computer Science* 117 (2017): 256-265.

[25] Altowayan, A. Aziz, and Ashraf Elnagar. "Improving Arabic sentiment analysis with sentiment-specific embeddings." In Big Data (Big Data), 2017 IEEE International Conference on, pp. 4314-4320. *IEEE*, 2017

[26] Alayba, Abdulaziz M., Vasile Palade, Matthew England, and Rahat Iqbal. "Improving Sentiment Analysis in Arabic Using Word Representation." arXiv preprint arXiv: 1803.00124(2018).

[27] Pagolu, Venkata Sasank, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi. "Sentiment analysis of Twitter data for predicting stock market movements." In Signal Processing, Communication, Power and Embedded System (SCOPES), 2016 International Conference on, pp. 1345-1350. *IEEE*, 2016.

[28] Sara Elshobaky and Noha Adly." A Platform For Sentiment Analysis on Arabic Social Media",2018, *International Conference on Data Science*

[29] Tang, Duyu, Furu Wei, Bing Qin, Nan Yang, Ting Liu, and Ming Zhou. "Sentiment embeddings with applications to sentiment analysis." *IEEE Transactions on Knowledge and Data Engineering* 28, no. 2 (2016): 496-509.

[30] Kharde, Vishal, and Prof Sonawane. "Sentiment analysis of twitter data: a survey of techniques." arXiv preprint arXiv: 1601.06971 (2016).

[31] Tang, Duyu, Bing Qin, and Ting Liu. "Deep learning for sentiment analysis: successful approaches and future challenges." Wiley Interdisciplinary Reviews: *Data Mining and Knowledge* Discovery 5, no. 6 (2015): 292-303.

[32] Yu, Liang-Chih, Jin Wang, K. Robert Lai, and Xuejie Zhang. "Refining word embeddings for sentiment analysis." In Proceedings of the *2017 Conference on Empirical Methods in Natural Language Processing*, pp. 534-539. 2017.