Bitlis Eren Üniversitesi Fen Bilimleri Dergisi
BİTLİS EREN UNIVERSITY JOURNAL OF SCIENCE
ISSN: 2147-3129/e-ISSN: 2147-3188
VOLUME: 12 NO: 2 PAGE: 412-417 YEAR: 2023
DOI:10.17798/bitlisfen.1229868

Sentiment Analyzing from Tweet Data's Using Bag of Words and Word2Vec

Yıldız AYDIN[1*]

[1]*Department of Computer Engineering, Erzincan Binali Yildirim University, Erzincan 24000, Turkey*
(ORCID: 0000-0002-3877-6782)

Keywords Natural language processing, Bag of Words, Word2Vec, sentiment analysis

Abstract
Twitter sentiment classification is an artificial approach for examining textual information and figuring out what people's public tweets from a variety of industries are experiencing or thinking. For instance, a large number of tweets containing hashtags are posted online every minute from one user to some other user in the commercial and political fields. It can be challenging for scientists to correctly comprehend the context in which specific tweet terms are used, necessitating a challenge in determining what is actually a positive or negative comment from the vast database of twitter data. The system's authenticity is violated by this issue and user dependability may be significantly diminished. In this study, twitter data sent to interpret movies were classified using various classifiers and feature methods. In this context, the IMDB database consisting of 50000 movie reviews was used. For the purpose of anticipating the sentimental tweets for categorization, a huge proportion of twitter data is analyzed. In the proposed method, bag of words and word2vec methods are given by combining them instead of giving them separately to the classifier. With both the suggested technique, the system's effectiveness is increased and the data that are empirically obtained from the real world situation may be distinguished well. With experimental efficiency of 90%, the suggested approach algorithms' output attempts to assess the reviews' tweets as well as be able to recognize movie reviews..

1. Introduction

The artificially intelligent technique of using natural language to communicate with an information system is known as natural language processing (NLP). Sentiment classification on social media platforms like Facebook and Twitter is one of the most popular uses of NLP. Chatbot, voice recognition, translation software, grammar verification, phrase scanning, information retrieval, and advertising pairing are the possibilities at the following level. This one will make it possible to create a twitter program that can comprehend linguistic forms. It is a component of data analytics that holds the effective algorithm of comprehending, extrapolating, and analyzing text data in the appropriate manner. Which enables one to process

enormous amounts of textual information, carry out several common operations, and provide answers to a specific set of issues for the previously mentioned future step NLP applications.

NLP and Twitter research are used while performing sentiment analysis and determining the individual emotions of tweets' necessary data. Sentiment classification is now being used to manage tweets' good, unfavorable, and neutral opinions. Sentiment analysis is commonly used in the spirit of community evaluations and poll answers, as well as medical billing and coding information that ranges from a marketing agency to civil administration. It may be

*Corresponding author: yciltas@erzincan.edu.tr
Received: 05.01.2023, Accepted: 09.05.2023

used to write textual information, photos, videos, voice recognition, and other types of information. Sentiment Classification uses comparable types of tweet data; after obtaining this information, it separates the input into individual words or phrases. This process is known as tokenization.

A movie review is similar to a language that expresses a view of the film and forecasts whether the general audience will have a favorable or unfavorable judgment. From all this, we may determine whether to see a certain film or not. A group of producers, actors, and others who participate in making films will be there. The community opinions made on specific cinematic sites determine a film's either success or failure. Website is essential tool for people since they allow them to express their thoughts on a film and submit customer reviews. Community evaluations are therefore gathered as a twitter data set, and after that, specific operations are carried out with this method to find the purchasing habits of customers and to foresee any upcoming releases in the film industry. For this purpose, sentiment analysis studies were conducted with the use of tweets expressing opinions about films in the literature. After receiving the twitter data and information from WordNet, Sahu and Ahuja [1] utilized SentiWordNet to determine the overall polarization of the film critic tweets. The mood of the phrases is included, and the ratings are divided into and positive. Also, there are objective scores in this method. The relative strength of it is determined by the measurement of each of those ratings. An artificial neural model called ConvLstm, built on the CNN and LSTM algorithms, was suggested by Hassan and Mahmood [2]. To lessen the waste of particular local knowledge and catch long-term interdependence in the series of words, they use LSTM as a replacement for CNN's pooling layers. The IMDB and Stanford Sentiment Treebank (SSTb) sentiment databases were used to test this concept. In order to compare their model to SVM and NB, Liao et al. [3] built a straightforward CNN method using word2vec using the data from twitter they acquired. CNN has therefore demonstrated to have an uses better accuracy when it comes to accuracy in comparison to other methods.

Sentiment analysis is now mostly concentrated on social networking, which includes sites like Facebook, Twitter, and IMDB [3], which is driving up demand for general opinions data and gathering it in textual form. Additionally, it is a difficult effort to anticipate the twitter data with an adequate understanding of the language for a movie critic. This study uses the IMDB database for sentiment analysis on English film reviews and identifies the reviews as either positive or negative applying classical machine classifiers. The primary objective of this research is to obtain more successful results by obtaining stronger features with the combination of bag of words and word2vec techniques. Also, NLP feature called stop keywords, which is used to filter frequently occurring tweet terms in the IMDB dataset, is also preprocessed in the proposed method. Due to the fact that these keywords will take up more room in our movie review information or they will fail to provide a coherent phrase for that information. This text content was used to eliminate certain terms for this cause. Word2vec and the bag of words (BoW) method were used in the feature extraction step, while support vector machine (SVM) and logistic regression methods were used in the classification step.

2. Material and Method

A paradigm for sentiment classification is suggested within that chapter. We must analyze the content to extract customer sentiment in an orderly fashion, despite the fact that we are aware that data from social media is not always clear, not always arranged appropriately, and sometimes even contains grammatical mistakes. As a result, we must pre-process the text in order to discern the text's mood. In order to determine whether a particular event is positive or negative, after the preprocessing step, the word is vectorized into numerical form and feature vectors are obtained. In the last step, classification is made using the obtained feature vectors. The methods used in the proposed approach are detailed in the following sub-sections. Figure 1 (the proposed framework for sentiment analysis) depicts all the processes.

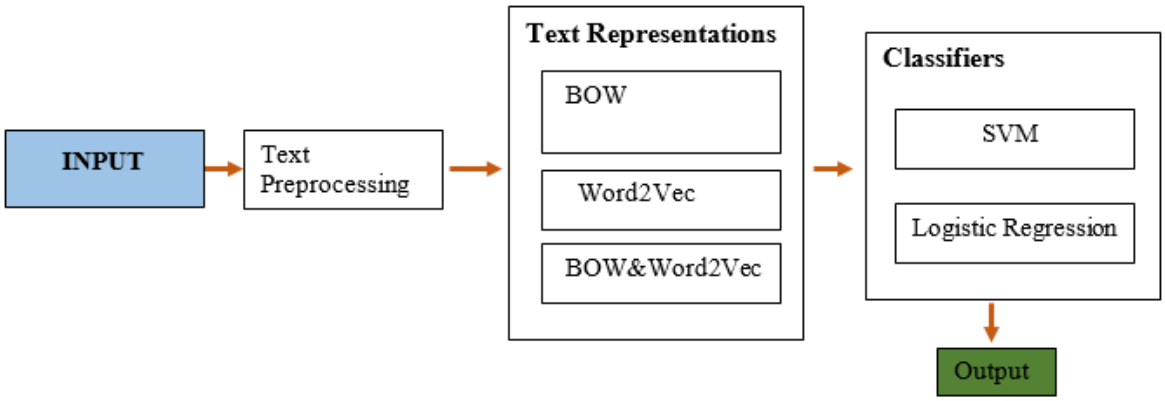

Figure 1. The proposed framework for sentiment analysis

1.1. Pre-processing

To provide a more distinctive representation of tweets data, the following actions have been taken

Cleaning: The goal of cleanliness is to eliminate unnecessary grammar and characters within textual data.

Stop Words Removal: Eliminating words such "the," "a," "an," "they," "she," and others that have no significance and do not sound right inside the content.

.

1.2. Text Representations

Bag of Words

Bag-of-Words (BoW) [4] presentations is widely used techniques in the field [5]. The i^{th} component of the matrix in the context of the BoW approach denotes the frequency with which the i^{th} vocabulary phrase appeared in the provided content. In other words, rather of utilizing a binary 1 or 0, frequency analysis is utilized to indicate the frequency of a certain word. Although these techniques are straightforward, the presentation does not keep the word's meaning. The linear combination is dense and have the same size as the vocabulary.

Word2Vec

Two linguistic methods are used by Word2vec [6]: the continuous bag-of-words (CBOW) method and the skip-gram method. These techniques use shallow neural network models to map the original word(s) to the targeted word(s).

Processing elements are then used to symbolize the phrases after the system has learned their weights throughout learning. In the CBOW approach, a given target word is produced from a collection of nearby terms. However, the skip-gram method operates precisely the opposite of the CBOW method. The Skip-gram method uses a single origin word as its input and seeks to identify a cluster of nearby words as its result.

Bag of Words & Word2Vec

The columns of the feature vectors obtained using the bag of words and Word2vec methods were combined and used in the Classifier. The same procedure was applied for each sample.

1.3. Classifiers

Support Vector Machines

Finding a hyper-plane that divided the training data set into two distinct groups is the basic idea behind this classification method [7]. The equations following may be used to make the following claim:

$$\boldsymbol{w^t x + b = 0} \tag{1}$$

wherein w: denotes the hyper-direction plane's. b depicts the location of the hyper-plane with regard to the origin.

Logistic Regression:

As a technique for classification model, we have employed logistic regression. The logistic regression method, which is one of the supervised classification methods, classifies the test sample using the logistic function [8].

3. Results And Discussion

During the testing stage, an Intel Core i7 CPU, 64-bit operating software, and 8GB RAM were used to run the system design in Python. The IMDB film critic database was used in the suggested approach. The IMDB database includes 50000 reviews in total, of which 25000 are labeled positive and 25000 are labeled negative.

The validity criteria are taken into account as the measuring performance of certain tests on databases for recall, precision, and F score in order to more clearly illustrate the method's analysis. Recall shows the ratio of positive samples correctly detected by the classifier to samples that are actually positively labeled (Equation 2).

$$\boldsymbol{recall}(\boldsymbol{r_c}) = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad (2)$$

Precision, on the other hand, shows the ratio of samples labeled positively by the classifier and samples correctly detected by the classifier (Equation 3).

$$\text{Precision}(\boldsymbol{p_r}) = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \quad (3)$$

F score is a metric obtained from precision and recall values, and the size of the F score shows the success of the application (Equation 4).

$$\boldsymbol{F1\ Score} = \mathbf{2} * \frac{\boldsymbol{Recall * Precision}}{\boldsymbol{Recall + Precision}} \quad (4)$$

In the experiments carried out within the scope of this manuscript, the feature vectors obtained from the word2vec, bag of words methods and the hybrid feature obtained from the combination of these two features were given to the SVM, logistic regression classifiers and their classifier performances were evaluated. The recall, precision and F1 values of the experiments are given in Table 1, and the confusion matrix graphics are given in Figure 2.

Tablo 1. Results of experiments

Features	Classifier	Precision	Recall	F-score	Accuracy
Bag of Words	SVM	0.89	0.87	0.88	0.88
	Logistic regression	0.88	0.87	0.87	0.87
Word2Vec	SVM	0.88	0.86	0.87	0.87
	Logistic regression	0.86	0.86	0.86	0.86
Proposed-Method	SVM	**0.90**	**0.89**	**0.89**	**0.89**
	Logistic regression	**0.89**	**0.88**	**0.89**	**0.89**

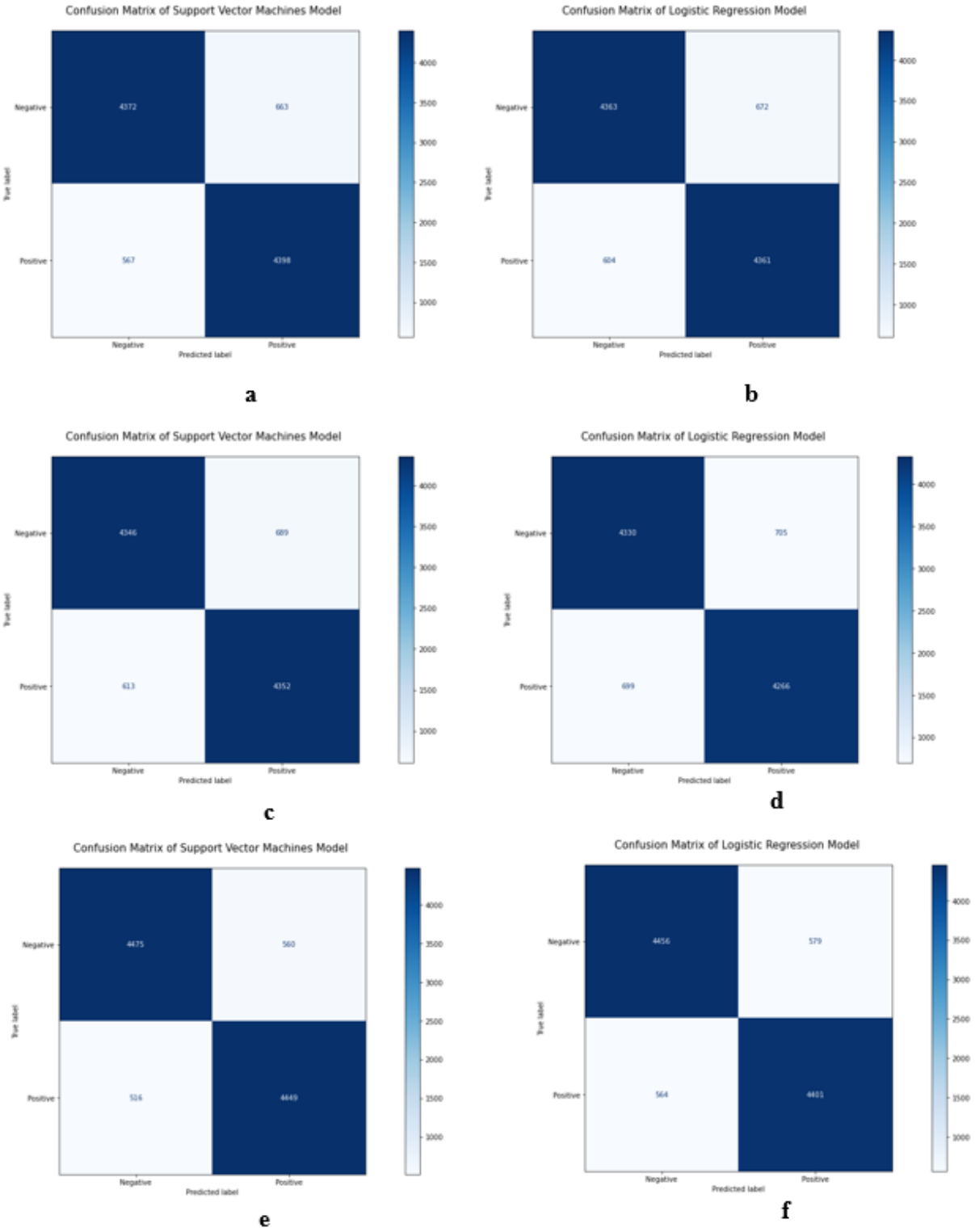

Figure 1. Confusion matrix of applications. a) BoW+SVM, b)BoW+Logistic Regression, c) Word2Vec+SVM, d) Word2Vec+Logistic Regression, e) BoW&Word2vec+SVM, f) BoW&Word2vec +Logistic Regression

As can be seen from Table 1 and Figure 2, using the features together instead of using them separately gave more successful results.

4. Conclusion and Suggestions

In this manuscript, it is proposed to use hybrid features to perform sentiment analysis from tweet data. Word2vec, bag of words and the vector obtained by combining these two methods were used as word representation techniques. As a classifier, the performance of the test results was evaluated by using SVM and logistic regression. In the experiments carried out using the IMDB dataset, 0.88 accuracy value was obtained with bag of words and SVM, 87 percent accuracy rate was obtained with Word2Vec and SVM, while the success rate in the proposed method was 0.89, which was the most successful method.

In addition, more successful results were obtained by combining these two methods rather than using a single bag of words or Word2Vec in other classifiers. In future studies, it is planned to perform sentiment analysis with the hybrid use of these word representation methods and deep learning methods.

Acknowledgment

This project did not receive any funding source

Conflict of Interest Statement

There is no conflict of interest between the authors.

Statement of Research and Publication Ethics

The study is complied with research and publication ethics

References

[1] T. P. Sahu and S. Ahuja, "Sentiment analysis of movie reviews: A study on feature selection and classification algorithms," *Int. Conf. Microelectron. Comput. Commun. MicroCom 2016*, 2016, doi: 10.1109/MicroCom.2016.7522583.

[2] A. Hassan and A. Mahmood, "Deep Learning approach for sentiment analysis of short texts," *2017 3rd Int. Conf. Control. Autom. Robot. ICCAR 2017*, pp. 705–710, 2017, doi: 10.1109/ICCAR.2017.7942788.

[3] S. Liao, J. Wang, R. Yu, K. Sato, and Z. Cheng, "CNN for situations understanding based on sentiment analysis of twitter data," *Procedia Comput. Sci.*, vol. 111, no. 2015, pp. 376–381, 2017, doi: 10.1016/j.procs.2017.06.037.

[4] J. R. Quınlan, "Programs for machine learning. Part II," in *Machine Learning*, vol. 7, pp. 135–240, 1994.

[5] J. D. Bodapati, N. Veeranjaneyulu, and S. Shaik, "Ingenierie des Systemes d ' Information Sentiment Analysis from Movie Reviews Using LSTMs," vol. 24, no. 1, pp. 125–129, 2019.

[6] T. Mikolov, K. Chen, G. S. Corrado, and J. A. Dean, "Computing numeric representations of words in a high-dimensional space," 2015.

[7] C. CORTES and V. VAPNIK, "Support-Vector Networks," *Mach. Leaming*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1109/64.163674.

[8] Y. Y. Liu, M. Yang, M. Ramsay, X. S. Li, and J. W. Coid, "A Comparison of Logistic Regression, Classification and Regression Tree, and Neural Networks Models in Predicting Violent Re-Offending," *J. Quant. Criminol.*, vol. 27, no. 4, pp. 547–573, 2011, doi: 10.1007/s10940-011-9137-7.