ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal
Regular Issue, Vol. 9 N. 3 (2020), 5-15
eISSN: 2255-2863
DOI: https://doi.org/10.14201/ADCAIJ202093515

Sentiment Analysis with Machine Learning Methods on Social Media

Muhammet Sinan Başarslan[a,b], Fatih Kayaalp[b]

[a] Computer Programming, Doğuş University, Istanbul, Turkey, 34775
[b] Department of Computer Engineering, Düzce University, Düzce, Turkey, 81620
mbasarslan@dogus.edu.tr, fatihkayaalp@duzce.edu.tr

KEYWORD

Sentiment analysis; Social Media; Python; Natural Language Processing.

ABSTRACT

Social media has become an important part of our everyday life due to the widespread use of the Internet. Of the social media services, Twitter is among the most used ones around the world. People share their opinions by writing tweets about numerous subjects, such as politics, sports, economy, etc. Millions of tweets per day create a huge dataset, which drew attention of the data scientists to focus on these data for sentiment analysis. The sentiment analysis focuses to identify the social media posts of users about a specific topic and categorize them as positive, negative or neutral. Thus, the study aims to investigate the effect of types of text representation on the performance of sentiment analysis. In this study, two datasets were used in the experiments. The first one is the user reviews about movies from the IMDB, which has been labeled by Kotzias, and the second one is the Twitter tweets, including the tweets of users about health topic in English in 2019, collected using the Twitter API. The Python programming language was used in the study both for implementing the classification models using the Naïve Bayes (NB), Support Vector Machines (SVM) and Artificial Neural Networks (ANN) algorithms, and for categorizing the sentiments as positive, negative and neutral. The feature extraction from the dataset was performed using Term Frequency-Inverse Document Frequency (TF-IDF) and Word2Vec (W2V) modeling techniques. The success percentages of the classification algorithms were compared at the end. According to the experimental results, Artificial Neural Network had the best accuracy performance in both datasets compared to the others.

1. Introduction

Thanks to the Internet, the developments in communication technologies have brought people closer together in recent years. The slow communication process of the past, using letters and telegraph, has now become an instant communication with the help of the Internet. Thanks to the social media, which is one of the application software emerged in line with the smartphone technology, communication environment now allows people to come into contact with everyone. This affected all people and institutions. For example, sharing a place, such as a movie theater, a store or a cafe, or expressing a positive or negative opinion about them affects everyone and the whole society in every field. People consider the social media as the main environment for communication. People share events, sports, film, personal feelings and thoughts that affect them through social media. This has transformed social media platforms into a large source of data, used by various entities ranging from businesses that want to promote or sell products, and scientific studies about people's feelings and ideas.

The fact that social media is an indispensable tool for people, and that they constantly express ideas about social, economic, health issues, and the products and brands has paved the way for sentiment analysis. In the sentiment analysis studies, sentiment expressions in the texts are predicted. The texts shared by people are examined in terms of their positivity, negativity or neutrality. Sentiment analysis allows a preliminary study on new products, new movies, etc. to be introduced by businesses.

The sentiment analysis process is performed on the data labeled as positive, negative or unbiased and sentiment estimation is carried out using various classification algorithms. Text preprocessing is performed via the text mining methods before the classification. To give an example of these processes, the symbols, punctuation in the text, and stems of the words, and the stop words are removed to create a list of terms, and the term frequencies and inverse document frequencies are used to create a vector space model. Sentiment analysis is performed by a classification process after obtaining the vector space model.

In their sentiment analysis study, Pang *et al.,* (Pang *et al*., 2002) have created a pre-classification vector space model on the movie comments in the Internet Movie Database archive, and conducted a sentiment analysis via classifying algorithms, such as Naïve Bayes (NB), Maximum Entropy and Support Vector Machine (SVM). Of the classification algorithms, the best performance has been obtained with SVM by 82.9% accuracy on the dataset using unigrams.

Another study conducted an emotion analysis using the SVM, NB classifiers, after obtaining the TF-IDF on the tweets sent during the 2012 Egyptian presidency election (Elghazaly *et al.,* 2016). According to the comparison made in their study, the NB method had the highest accuracy and lowest error rate.

Hamoud et al., (Hamoud *et al.,* 2018) have used the Bag of Words (BOW), TF and TF-IDF on the Twitter data for the classification of political tweets. Of the classification algorithms, they have used SVM and NB. According to the results, BOW-enabled SVM provides the highest accuracy and F-measure.

Nikfarjam et al. (Nikfarjam *et al.,* 2015) have conducted a sentiment analysis on the Twitter using the comments of patients about the side effects of drugs. At the end of their study, they stated that the SVM algorithm performs better by 82.1% compared to the other methods.

There are many studies conducted on Turkish datasets. In their study, *Nizam et al.,* (Nizam and Akın, 2014) have investigated whether the distribution of the data in the classes had an effect on the success rate of the classification algorithm and found that the data distribution is of importance for the

success rate. In their study on the food industry data, they obtained an accuracy rate of 72.33% using the SVM algorithm.

In another study, *Türkmen et al.* (Türkmen *et al.,* 2014) used various classifying algorithms, such as the decision tree, k-nearest neighbor, NB and SVM to calculate sentiment polarity (poles) on Turkish movie comments. They achieved the best result by the SVM.

In their study on movie reviews, *Kaynar et al.* (Kaynar *et al.,* 2018) used the Naïve Bayes, Multilayered Artificial Neural Network, and Support Vector Machine Algorithms. They used TF-IDF for the feature extraction. The support vector machine has yielded better results than other algorithms.

In their sentiment analysis study, *Huq et al.* (Huq *et al.,* 2017) used the SVM and k-NN machine learning algorithms on the Twitter data, and obtained the normal tracking accuracy values between 58.39% and 79.99% on the datasets obtained after the feature extraction by the n-grams.

Amolik et al. (Amolik *et al.,* 2016) proposed sentiment analysis, and they accurately classified tweets by using the Feature-Vector and classifiers like NB and SVM. In spite of the lower recall and accuracy, NB has better precision compared to SVM. However, SVM gives better result when it comes to accuracy.

*Symeonidis et al. (*Symeonidis *et al.,* 2018) used Linear SVC, Bernoulli Naïve Bayes, Logistic Regression and Convolutional Neural Networks, which are four popular machine learning algorithms. The author achieved the best results by the CNN.

Rana and Singh (Rana and Singh, 2016) carried out emotional analysis on the texts in various categories using algorithms such as Naïve Bayes, Linear SVM and Synthetic words. In the experimental results, Linear SVM was found to provide the best accuracy, followed by the Synthetic words approach.

In this study, tweets tagged in health topic were collected from the Twitter in 2019. After pre-processing the tweets by text mining, a vector space model was obtained by using the term frequencies and inverse document frequencies; and then the sentiment analysis was performed by using the ANN, SVM, and NB classifier algorithms. Under the second chapter, the Twitter dataset, and the text mining and classification algorithms used in the study are presented. The experimental results are given in the third chapter, and the final chapter presents the discussions and conclusions.

1.1. Contribution of Study

As shown in the related studies given above, the machine learning classifier algorithms such as SVM, ANN and NB are popular and have good performance in sentiment analysis studies. As a contribution, this study evaluates the performance of these algorithms in comparison with the traditional frequency-based text representation (TF-IDF) and prediction-based text representation (W2V) methods.

According to the results of the experiments on the datasets such as IMDB, Yelp and the tweets collected and tagged according to the sentiments by the researchers, the model created by the W2V and ANN had better performance compared to the others.

2. Method

This section gives information about the Twitter and the dataset, text mining, and classification algorithms used in the study. The flowchart of the study is shown in Fig.1.

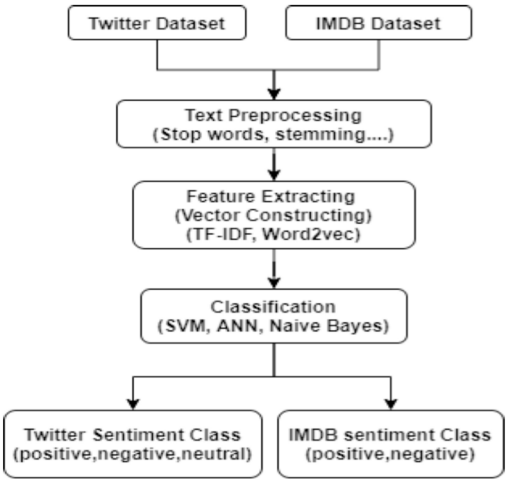

Fig.1. The Flowchart of the Study.

2.1. Datasets

Twitter is a microblogging site where Jack Dorsey can share text, pictures, or videos instantly with 280 character restrictions. You can also follow other accounts, like tweets sent from those accounts or retweet them again (Rogers, 2014).

For this study, 4500 health-related twitter data were collected using the Twitter (Application Programming Interface) API. The preprocessing and sentiment scores of these data were carried out by a Python program. Of the tweets collected and labeled, 1680 were neutral, 1220 were positive, and 1600 were negative. Table 1 shows the attributes of the tweets collected by the API via Python. In addition to the study on the data collected from the Twitter, the same models were also applied to the dataset consisting of 500 positive and 500 negative opinions collected by Kotzias *et al.* (Kotzias *et al.,* 2015) from the IMDB movie reviews, given in Table 2.

The neutral-tagged tweets from the data collected from the Twitter were the drug ads, and their attribute information is given in Table 1. Tweets marked as negative seem to belong to those with various diseases. On the other hand, the positive ones are the tweets indicating that the diseases such as cancer have successfully treated.

Table 1: Twitter dataset

Dataset Attribute	Explanation of Attribute
id	Order of tweet dataframe
text	tweet
created_at	Date and time the Tweet was posted
retweeted	Tweet rerun status (bool)

Dataset Attribute	Explanation of Attribute
retweet_count	Number of retweets
user_screen_name	Username
user_followers_ count	Number of followers
user_location	Followers location
hashtags	Tweet tag
sentiment_score	Sentiment score
sentiment_class	positive, negative, neutral

Table 2: The IMDB dataset of Kotzias

Dataset Attribute	Explanation of Attribute
text	Reviews from imdb
sentiment class	positive, negative

2.2. Text Mining

Prior to the classification, 4500 Twitter records were preprocessed. The symbols and punctuation marks in the comments were removed, the characters were converted to the lower case, the root of each word was found, and the stop words, such as "@username" were omitted. The Python NLTK library was used for these operations.

The term frequency (TF) and inverse document frequency (IDF) were used for the feature extraction. TF is the frequency of occurrence of the terms in the documents. The IDF is used to normalize the term frequencies, although the terms in the text often have no distinguishing significance.

2.2.1. TF-IDF (Term Frequency-Inverse Document Frequency)

The TF is the method used to calculate term weights in a document as seen in Eq. (1). The IDF finds out the number of words in more than one document and determines whether the word is a term or not (Stop Words). For this purpose, the absolute value of the logarithm of the number of documents that contains the term must be divided by the number of documents, as shown in Eq. (2) (Sjögren *et al.,* 2020)

TF-IDF score form term i in document j=TF(i,j)* IDF(i)

$$TF(i,j) = \frac{Term\ i\ frequency\ in\ document\ j}{Total\ \ words\ in\ document\ j} \quad (1)$$

$$IDF(i) = log\left(\frac{Total\ documents}{documents\ with\ term\ i}\right) \quad (2)$$

t = Term, j =Document

2.2.2. Word2vec

Word2vec is an unsupervised natural language processing tool that uses the artificial neural network structure developed by Mikolov *et al.* (Mikolov *et al.,* 2013). The tool takes a text as the input and represents each word in the text in a vector. Basically, word2vec clears the semantically similar words in close coordinates. Two different learning architectures, continuous bag of words (CBOW) and skip-gram (SG), were used to find the word coordinates. In the CBOW architecture, neighboring words (words to the right and left) of a word within a certain window size are examined, and the word estimation is performed through the neighboring words. In the skip-gram architecture, neighboring words are estimated by looking at the target word in the opposite way.

2.3. Classification Algorithms

Support Vector Machine, NB and Artificial Neural Network classifier algorithms were used in the study. These three algorithms are discussed in detail in this section.

2.3.1. Naïve Bayes

The Naïve Bayes (NB) algorithm was named after the English mathematician Thomas Bayes. Bayesian algorithms are among the statistical classification techniques and are based on the statistical Bayesian theorem. Bayes classifier is a predictive model, which is easier to apply.

Let X = x_1, x_2, x_3, … … x_n, is the sample set, and C_1, C_2, C_3, …C_m is the class set. The sample to be classified,

$$P(X|C_i) = \frac{P(X|C_i)P(P(C_i)}{P(X)} \quad (3)$$

The probability is calculated as shown in Eq. (3). The data sample with the highest probability, calculated for each class, belongs to that class (Elmas, 2015).

2.3.2. Support Vector Machine

The SVM can be defined as a vector space-based data mining method that finds a decision boundary between the two classes farthest from a random point on the training data (Song *et al.,* 2002). An interesting feature of the SVM is its structural risk minimization in statistical learning theory. One of the main assumptions of the SVM.

2.3.3. Artificial Neural Network

Artificial neural networks are parallel and distributed information processing structures that are inspired by the human brain, and are composed of processing elements, each of which has its own

memory, connected to each other via weighted connections. Artificial neural networks, in other words, are the computer programs that mimic biological neural networks (Kayikci and Akyazi, 2012).

The structure of the artificial neural networks consists of three components: the neuron (artificial nerve cell), the connections, and the learning algorithm. The neuron is the basic processing element of an artificial neural network. The neurons in the network receive one or more input according to the factors that affect the problem, and output the number of results expected from the problem. The connection of neurons forms an artificial neural network (Fig. 2.). In a general artificial neural network system, neurons come together in the same direction to form layers (Harrington, 2012).

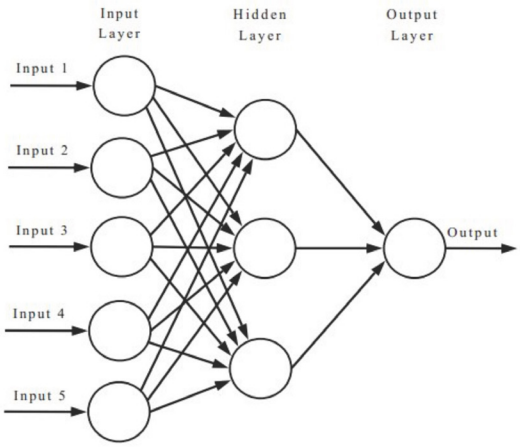

Fig.2. Artificial Neural Network Architecture.

2.4. Performance Criteria

In this study, confusion matrix was used to evaluate the models developed with classification algorithms (Wright *et al.,* 2020). For the performance evaluation, four statistical measures were used, including the accuracy (ACC), sensitivity (SENS), specificity (SPEC), and F-measure (F). SENS represents the probability of correctly identifying the True Positive (TP) class, Y means 'Yes', while specificity represents the probability of correctly identifying the True Negative (TN) class, and here Y means 'No'. If the model predicts a class as negative, while the actual class is positive, we define it as False Negative (FN). On the contrary, if the model predicts a class as positive, while the actual class is negative, we define it as False Positive (FP). Overall, the accuracy measures the probability of detecting the true class. Eq. (4-7) is the harmonic mean of the precision and recall, where an F measure reaches its best value at 1 (perfect SPEC and SENS) and the worst at 0 (Xiao *et al.,* 2020).

The accuracy value is shown in Eq. (4).

$$Accuracy = \frac{T_P + T_N}{T_P + F_N + F_P + F_N} \tag{4}$$

Sensitivity value is shown in Eq. (5).

$$Sensitivity = \frac{T_P}{T_P + F_N} \tag{5}$$

Precision value is shown in Eq. (6).

$$Precision = \frac{T_P}{T_P + F_P} \tag{6}$$

F-measure value is shown in Eq. (7).

$$F\text{-}measure = \frac{2 * Precision * Sensitivity}{Precision + Sensitivity} \tag{7}$$

The data sets are divided into two as test and training in order to establish the model with classifier algorithms and then to evaluate the performance of these models. In the study, it was preferred to work with k-fold cross validation. The cross-validation method and test and training cluster separation used in the study are shown in Fig 3.

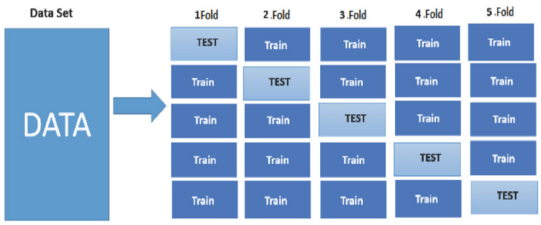

Fig. 3. Cross-validation method, and test and training set separation.

3. Experimental Result

In the study, 4500 tweets were used. The NB, SVM, ANN was used for the classification in the sentiment analysis. After the text pre-processing and vector space modeling, 5-fold cross-validation was used for separating training and test data sets. Then, the performance of these three algorithms was evaluated by using the accuracy (ACC), precision (PRE), sensitivity (SENS) and F-measure (F) parameters. The results are shown in Table 4.

In addition, the performance results of the given classification algorithms on IMDB dataset, including polarities labeled by Kotzias, are given in Table 5.

Table 4. Results with TF-IDF on Twitter and IMDB datasets

Datasets	Algorithm	ACC	PREC	SENS	F
IMDB	**NB**	0.82	0.82	0.83	0.82
	SVM	0.83	0.84	0.84	0.84
	ANN	**0,89**	**0,88**	**0,88**	**0,89**
Twitter	**NB**	0,72	0.73	0.72	0,76
	SVM	0,82	0.83	0.82	0,81
	ANN	**0,86**	**0,87**	**0,84**	**0,85**

Table 5. Results with W2V on the Twitter and IMDB datasets

Datasets	Algorithm	ACC	PREC	SENS	F
IMDB Dataset	**NB**	0.83	0.84	0.85	0.84
	SVM	0.84	0.84	0.86	0.85
	ANN	**0,90**	**0,91**	**0,90**	**0,96**
Twitter Dataset	**NB**	0,72	0.76	0.76	0,77
	SVM	0,84	0.85	0.84	0,82
	ANN	**0,87**	**0,88**	**0,86**	**0,86**

In order to verify the performance results of the classifiers for the algorithms on the labeled IMDB dataset, the same algorithms were applied to the Twitter dataset after the TF-IDF and W2V methods used for the vector modeling. After analyzing the two datasets in Table 3 together following the TF-IDF word vector process, the Twitter and IMDB datasets were analyzed, and the performance values were found to be approximately the same. The ANN gave the best performance among others in both datasets. In addition, the NB gave the worst performance among others in both datasets. Table 4 presents better performance results compared to that of Table 3. However, in the W2V method shown in Table 4, the performance of classification algorithms was found to increase.

4. Conclusion

In the study, the success of the classifiers was investigated on the two datasets, one from IMDB and one from Twitter. The classifier algorithms used in the experiments were ANN, SVM and Naïve Bayes. The results were obtained by dividing the datasets into the training and test data sets by the 5-fold cross validation after the text pre-processing and vector space modeling for the sentiment analysis classification. The ACC, PREC, SENS and F-measure were used to evaluate the performance.

The classification experiments were performed on the Twitter dataset first. In addition, after obtaining the classification results, the same classification processes were applied to the IMDB dataset, labeled by Kotzias. Then, the performance results of the two experiments were compared to check whether the second results validate the first ones. After analyzing all the results of both experiments, it is clearly seen that the performances of the classification algorithms confirm each other. The performance values of the algorithms also had the same success rates. ANN had the best performance on all performance criteria. On the other hand, NB had the worst performance. In the future studies, it is aimed to use more advanced artificial neural network models for classification.

In order to compare sentiment analysis performance between two languages, performing sentiment analysis on the Turkish tweets in addition to the English is planned for future work. In addition, sentiment analysis will be carried out on the data taken from different websites and social media sites, other than the Twitter, where people share their opinions. Moreover, in the future studies, classifier models will be created with deep learning algorithms, following the common word embedding methods, such as Bert for displaying text.

5. References

Amolik, A., Jivane, N., Bhandari, M., and Venkatesan, M., 2016. Twitter sentiment analysis of movie reviews using machine learning techniques. International Journal of Engineering and Technology, 7(6): 1-7.

Elghazaly, T. Mahmoud, A. Hefny, H. A., 2016. Political sentiment analysis using twitter data. In: Proceedings of the International Conference on Internet of things and Cloud Computin,1-5.

Elmas, Ç., 2003. Yapay Sinir Ağları (Kuram, Mimari, Eğitim, Uygulama). Ankara: Seçkin Yayıncılık.

Harrington, P.,2012. Machine learning in action. Shelter Island, NY: Manning Publications Co.

Hamoud, A. A., Alwehaibi, A., Roy, K., and Bikdash, M. 2018. Classifying political tweets using Naïve Bayes and support vector machines. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems(736-744). Springer, Cham.

Huq, M. R., Ali, A., and Rahman, A., 2017. Sentiment analysis on Twitter data using KNN and SVM. (IJACSA) International Journal of Advanced Computer Science and Applications, 8(6): 19-25.

Kayikci, S., Akyazi, E., 2018. Classification of Open Directory Web Pages Using Artificial Neural Networks. International Journal of Scientific and Technological Research, 2422-8702

Kaynar, O., Görmez, Y., Yıldız, M., and Albayrak, A., 2016 .Makine öğrenmesi yöntemleri ile Duygu Analizi. In International Artificial Intelligence and Data Processing Symposium (IDAP'16), 17-18.

Kotzias, D., Denil, M., De Freitas, N., and Smyth, P. 2015. From group to individual labels using deep features. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 597-606.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J., 2013. Distributed compositionality. Advances in Neural Information Processing Systems. 26: 3111-3119.

Nikfarjam, A, Sarker, A, O'Connor, K, Ginn, R, and Gonzalez, G., 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features, Journal of the American Medical Informatics Association, 22(3): 671-681

Nizam, H, Akın, S. S.,2014. Sosyal medyada makine öğrenmesi ile duygu analizinde dengeli ve dengesiz veri setlerinin performanslarının karşılaştırılması. XIX. Türkiye'de İnternet Konferansı.

Pang, B., Lee, L., and Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. arXiv preprint cs/0205070.

Rana, S. and Singh, A.,2016. Comparative analysis of sentiment orientation using SVM and Naïve Bayes techniques, 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), Dehradun, pages 106-111, doi: 10.1109/NGCT.2016.7877399.

Rogers, R., 2014. Debanalising Twitter. Twitter and Society, New York, NY, ix-xxxviii.

Sjögren, R., Stridh, K., Skotare, T., and Trygg, J., 2020. Multivariate patent analysis–Using chemometrics to analyze collections of chemical and pharmaceutical patents. Journal of Chemometrics, 34(1): e3041.

Song, O., Hu, W., and Xie, W, 2002. Robust Support Vector Machine with Bullet Hole Image Classification, IEEE Transactions on Systems, Man and Cybernetics – Part C: Applications and Rewiews, 32(4): 440-448.

Symeonidis S, Effrosynidis D., and Arampatzis A., 2002. A comparative evaluation of pre-processing techniques and their interactions for Twitter sentiment analysis. Expert System Applications, 110:298-310.

Türkmen, A. C. Cemgil, A. T., 2014. Political interest and tendency prediction from microblog data. In: 22nd Signal Processing and Communications Applications Conference (SIU). IEEE, 1327-1330

Wright, G., Rodriguez, A., Li, J., Clark, P. L., Milenković, T., and Emrich, S. J., 2020. Analysis of computational codon usage models and their association with translationally slow codons. PloS one, 15(4): e0232003.

Xiao, C., Xia, W., and Jiang, J., 2020. Stock price forecast based on combined model of ARI-MA-LS-SVM. Neural Computing and Applications, 1-10.