J Inno Sci Eng,2021, 5(2):173-182
http://doi.org/tr/10.38088/jise.935963

Research Article

Keyword-based Sentiment Analysis of Covid-19 Related Tweets

Mustafa Ozgür Cingiz [1*] Ece Celiktas[2]

[1*,2] *Department of Computer Engineering, Bursa Technical University, 16310 Bursa, TURKEY*

Cite this paper as:
Cingiz, O. M. and Celiktaş, E. (2021). *Keyword-based Sentiment Analysis of Covid-19 Related Tweets*.Journal of Innovative Science and Engineering.5(2):173-182

*Corresponding author: Mustafa Ozgür Cingiz
E-mail: mustafa.cingiz@btu.edu.tr

Received Date:11/05/2021
Accepted Date:09/10/2021

© Copyright 2021 by Bursa Technical University. Available online at http://jise.btu.edu.tr/

The works published in Journal of Innovative Science and Engineering (JISE) are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Abstract

With the emergence of Web 2.0, internet users share their feelings, thoughts and ideas with other people using social networks. Understanding people's thought analysis is important for examining marketing and user feedback in social networks. For this reason, sentiment analysis on social networks with machine learning algorithms is a popular field of study. Our study is based on the sentiment analysis of people against the new coronavirus, which affects the world. People can have different moods due to pandemia. The governance of mental issues must be observed to manage the pandemic time period more successfully. In this article, we retrieved 387,953 tweets due to the ten most frequently used COVID-19 related keywords. The most frequently used keywords about COVID-19 which enable to obtain and assess the reaction of Twitter users are investigated. Even if COVID-19 is a health issue and tweets about COVID-19 is expected to contain negative content, we found positive, negative and neutral tweets to analyze texts using sentiment analysis and machine learning approaches. We applied four classifiers like logistic regression, multinomial naive Bayes, support vector machines and decision tree. These classifiers are well studied and utilized in many studies which we mentioned in our study. The performance of the support vector machine, decision tree and logistic regression classifiers are close to each other. The lowest F-score is obtained from multinominal naive Bayes classifier. The classification results for each negative, neutral and positive class were compared separately in our study.

Keywords: Social Networks, Microblogs, Sentiment Analysis, Covid-19, Classification.

.

1. Introduction

Twitter has become a social networking platform where people can instantly share their thoughts. People can share many situations they have experienced and encountered in their lives, or their thoughts on a subject, positively, negatively or objectively. Twitter has been doing a lot of research for guidance. Users are offered the opportunity to share emotions, thoughts and ideas (Tweet) limited to 280 characters on Twitter. The Twitter API allows us to analyze the responses of a particular community on the topic it addresses. Sentiment analysis has gained great importance in recent years due to market research and commercial use. The usage of machine learning algorithms enables researchers to model the opinions of users.

COVID-19 has been declared a global pandemic in the world and continues its effect. Twitter users ignore the importance of COVID-19 in the early stage even though general thoughts of users in Twitter reflect anxiety. In this study, English tweets containing COVID-19 content on the Twitter platform were handled and analyzed. During the pandemic process, the tweets shared by Twitter users on the issue and reacted to were collected and their opinions were analyzed.

There were many studies have been published about sentiment analysis using Twitter data, inferred or public datasets. Barkur [1] examined tweets of Indians during the first days of pandemia. The study intends to understand the mental health of people via examining tweets which are classified as positive, negative via word-cloud analysis. Chakraborty [2] used corona, covid, sarscov2, covid19 and coronavirus keywords to infer COVID-19 related tweets. They used different classifiers to make machine learning models and logistic regression and the tf-IDF model outperformed in the study. Alamoodi [3] surveyed lexicon, machine learning and hybrid techniques about sentiment analysis for COVID-19. Rustam [4] also utilized five different classifiers and tf-IDF, a bag of words patterns in the sentiment analysis of COVID-19 tweets. Sentiment analysis is also popular in different study areas. Ayan investigated whether the content is Islamophobic by analyzing the sentiment of tweets over social networks using the machine learning method using a distributed computing infrastructure [5]. By using naïve Bayes and support vector machines machine learning algorithms by İlhan, sentiment analysis was performed on Twitter data where users shared their status updates on Twitter [6]. Akın presented a sentiment analysis model using a dictionary-based analysis of positive and negative emotions [7]. Uslu applied the Dictionary-based support vector machines, naïve Bayes, logistic regression and decision tree approaches on the Turkish dataset in order to examine the emotions on user comments on movie websites [8]. Neethu and Rajasree aimed to analyze sentiment with the machine learning method. In their study, naïve Bayes, support vector machines have compared the successful performances of classifiers using collective learning approaches [9]. Saha aimed to analyze the product quality of companies. They classified the tweets with positive and negative content about the products using naïve Bayes and support vector machines [10]. Gautam [11] discussed a machine learning algorithm with semantic analysis for classifying the sentence and product reviews based on Twitter data. Wongkar [12] investigated the sentiment analysis of the public towards the Republic of Indonesia's presidential candidates. They compared SVM and KNN algorithms in their study. Mandloi [13] utilized different machine learning techniques of data analysis of twitter that are discussed naïve Bayes, SVM and maximum entropy method in their study. El Rahman [14] analyzed opinions on several social media sites. In their study, they compared machine learning methods such as naïve Bayes classification method, support vector machine and maximum entropy classification methods. Al Shammari [15] investigated the problem of real-time Twitter sentiment analysis which relies on providing a graphical representation of tweets categories (positive, negative, and neutral) opinions to help such as companies and agencies to focus on users' opinions of their products.

We created our own dataset for this study, by collecting tweets from Twitter users between March and May 2020. The time interval of tweet retrieval is important due to the pandemic period which indicates the beginning of the pandemia. The obtained tweets were taken from the first times of the coronavirus pandemic, so it allowed us to analyze the first feelings and thoughts of people about the pandemic. The similar studies utilized fewer keywords to retrieve tweets about COVID-19 but we determined keywords due to their usage and hashtag analysis. COVID-19 related tweets are collected by ten frequent keywords. We present materials and methods in the second part, results and discussion in the third part and conclusion in the last part of our study.

2. Materials and Methods

In our study, 10 subject tags were used to obtain tweets from Twitter, these keywords are: #Corona, #Covid, #Covid19, #CoronaVirus, #Wuhan, #Covid2019, #Corona, #CoronaVirusUpdate, Covid-19, Covid-2019. Twitter Stream API provides the opportunity to instantly collect tweets. Tweepy library dataset containing user tweets associated with COVID-19 was created [16]. NLTK library was used in Python software to perform pre-processing steps [17]. Positive, negative or neutral (not expressing opinion) information, which is the sentiment information of the contents, was obtained using the TextBlob library [18].

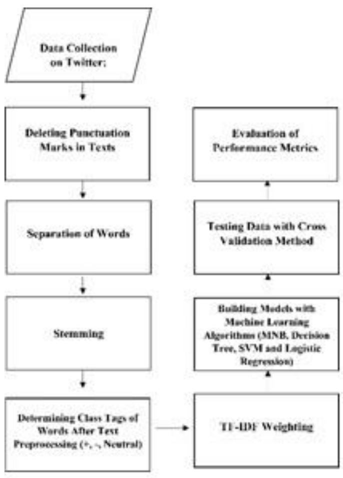

Figure 1. The Proposed System Model

In the classification and evaluation, machine learning algorithms were applied using the Sklearn library [19]. In this study, the steps of the proposed system model are shown in Figure 1 and details of the steps are given as subsections in the second part of the study.

2.1. Data Sets

387,953 tweets were retrieved as the dataset. The fetching date of these tweets is between March 2020 and May 2020. A total of 3 class labels, positive, negative and neutral were used in the dataset. It consists of a total of 387,953 data; 155,305 tweets in the positive class, 60,610 tweets in the negative class and 172,038 tweets in the neutral class. The dataset sample is presented in in Table 1.

Table 1. Samples and Classes from the Data Set

	Text(tweet)	Sentiment Class
1.	@qureshik74: My brave nephew stated his duty in the #Corona ward at Benazir Bhutto Hospital	Positive
2.	@ITSJEREMYSCOTT Irresponsible behavior will make you pay heavy... you destroyed others too #Corona	Negative
3.	One thing everyone talking about is #Covid_19	Neutral

2.2. Data Pre-processing

It is aimed to reveal the features of the text by cleaning it in the text pre-processing stage. These steps are as follows; Conversion of capital letters to lowercase letters, removal of punctuation marks and ineffective words, deletion of excess spaces and, finally, the process of retrieval of stems of words. Examples of the data set where text pre-processing steps are applied are shown in Table 2.

Table 2. Samples from the Pre-processed Data Set

	Text(tweet)
1.	['brave','nephew','stated','duty','corona','ward','benazir','bhutto','hospital']
2.	['irresponsible','behavior','make','pay','heavy','destroyed','others','corona']
3.	['one','thing','everyone','talking','covid']

2.3. Determining Class Labels

In this section, the classification of the pre-processed texts was carried out with TextBlob. Textblob is an NLP library based on Python language that is used for text processing. The inferred tweets of our study are labelled as positive, negative or neutral, based on sentiment analysis via Textblob. Table 3 shows some examples from the dataset that are labelled with sentiment analysis tags of Textblob.

Table 3. Dataset samples for classification

	Text(tweet)	Sentiment class
1.	['brave','nephew','stated','duty','corona','ward','benazir','bhutto','hospital']	Positive
2.	['irresponsible','behavior','make','pay','heavy','destroyed','others','corona']	Negative
3.	['one','thing','everyone','talking','covid']	Neutral

2.4. Term Weighting TF-IDF

Text documents are represented as vectors in the vector space model. Each dimension of the vector corresponds to words in the word space model. If a term is included in the text, its weight in the vector cannot be zero. There are many methods of expressing the weight of terms in the text. In this study, the frequently used TF-IDF (TF, term frequency-IDF, reverse document frequency) weighting method was utilized. TF is the frequency of a term in the document and indicates how many times the word occurs in the document.

The mathematical expression of the term frequency (TF) is given in equation (1). f_d (t) represents the number of occurrences of term tin document d the maximum number of term occurrence is used for normalization of term frequency.

$$tf(t,d) = \frac{f_d(t)}{\max_{\omega \in d} f_d(\omega)} \tag{1}$$

Reverse document frequency (IDF) shows how often the word occurs in all documents. The mathematical expression of the inverse document calculation is given in equation (2). |D| show number of documents and divisor part indicates the number of documents that term t occur.

$$idf(t,D) = ln\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right) \tag{2}$$

In TF-IDF weighting, it is obtained from the product of TF and IDF values, and the corresponding mathematical expression is given in equation (3).

$$tf.idf(t,d,D) = tf(t,d) \cdot idf(t,D) \tag{3}$$

A word has strong weight when it has high term frequency and low reverse document frequency according to Equation 3.

2.5. Classification

Multinominal Naïve Bayes, decision tree, support vector machines (SVM), logistic regression classification algorithms are applied on the retrieved dataset using the scikit-learn Python library. The classic Naïve Bayes classifier is a machine learning model. The classifier is based on the Bayes theorem. In this study, multinomial Naïve Bayes, which is a Naïve Bayes classifier type, was used. Multinomial Naïve Bayes is used especially for document classification problems. The logistic regression method is a method that provides flexibility in the classification of multivariate data. It performs classification operations by determining the probability value that predicts the logistic regression dependent variable and supports vector machine classifier basically performs classification using boundary data close to the decision-making axes. Decision boundaries, also known as the hyperplane, are determined. Support vector machines linear kernel function is used in our study. This method provides better scaling for datasets with a large number of samples. A decision tree classifier is a classification method that creates a tree structure model consisting of decision nodes and leaf nodes according to feature and target. The decision tree algorithm divides the data set into small pieces. A decision node can contain one or more branches. The first node is called a root node. A decision tree can consist of both categorical and numerical data. In this study, the parameter of the decision tree classifier was chosen as entropy. Linear SVC was used when calculating the support vector machine classifier.

2.5.1 Performance Metrics

The performance of the classifiers was evaluated by using the 10-fold cross-validation method and the k value is determined 10 in our study. Accuracy, precision, recall and F1-score criteria were used as classification performance evaluation metrics. True positive (TP) indicates a classifier predicts the positive class label of the instance correctly. If a classifier determines a positive class for a negative class instance then it is represented as false positive (FP) in the confusion matrix. If a classifier determines a negative class for a positive class instance then it is represented as a false negative (FN). The last metric of the confusion matrix is true negative (TN) which occurs when a classifier specifies a negative class label for an actual negative class instance.

Accuracy: This statistic is the correct prediction rate. The accurate prediction rate is divided by the total number of test observations of the test observations whose labels are correctly predicted. The mathematical expression of the accuracy is given in Equation 4.

$$\frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

Recall: It is how accurately we estimate from all positive classes. The mathematical expression of the recall is given in Equation 5. The recall represents how accurately a classifier specifies actual positive classes of instances. Its calculation is given in Equation 5.

$$\frac{TP}{TP+FN} \tag{5}$$

Precision: Precision represents how accurately a classifier specifies the ratio of positive labelled instances, which are found as correctly, to all positive predicted labelled instances. Its calculation is given in Equation 6.

$$\frac{TP}{TP+FP} \tag{6}$$

F1-Score: It is the harmonic mean of the Precision and Recall values. The mathematical expression of the F1-Score is given in equation (7).

$$\frac{2xRecallxPrecision}{Recall+Precision} \tag{7}$$

ROC (Receiver Operating Characteristic) Curve: ROC curve is a plot whose x-axis presents a false positive rate and the y-axis displays true positive rates. The false-positive rate indicates FP/ (FP+ TN) and the true positive rate shows recall value. The area under the ROC curve approximates to one means obtaining better classification results.

3. Results and Discussion

The true positive(TP), true negative(TN), false positive(FP), and false-negative (FN) values obtained for each classifier algorithm in this study are shown in Table 4.

Table 4. TP, TN, FP and FN values for each classifier

Classifier	**Class**	**TP**	**TN**	**FP**	**FN**
Multinominal Naive Bayes	Negative	822	32740	5192	5
	Neutral	15215	19869	1698	1946
	Positive	15218	17318	462	5873
Decision Tree	Negative	5469	32316	540	571
	Neutral	17167	21418	207	131
	Positive	14787	22705	629	567
SVM	Negative	5779	32565	253	209
	Neutral	17045	21473	105	173
	Positive	15388	23046	184	260
Logistic Regression	Negative	5447	32525	569	254
	Neutral	17051	21024	561	151
	Positive	15083	22886	448	380

The median values of TP, TN, FP and FN of 10-fold cross-validation are given in Table 4. Four classifiers performance results can be measured in each class. Multinominal naive Bayes classifier presents lowest TP values in negative tweets. It predicts negative tweets as positive due to TP and FP values of multinominal naive Bayes. TP values of SVM on negative and positive classes is higher than TP values of other classifiers. FN values of the Decision tree is also higher than FN values of SVM and logistic regression. The highest FN values are also obtained from multinominal naive Bayes. The lowest FP and FN values are obtained using SVM. We created a supplementary file that contains all TP, FP, FN and TN values for all classifiers on three classes using 10-fold cross-validation.

All TP, FN, FP, and FN values are summed to calculate more robust classification performance metrics. Accuracy, precision, recall and F- score values are presented in Table 5.

Table 5. Performance Metrics of Classifiers

	Accuracy	Precision	Recall	F1-Score
Multinomial Naive Bayes	0.80	0.87	0.66	0.65
Logistic Regression	0.96	0.95	0.95	0.96
Support Vector Machine	0.98	0.98	0.97	0.97
Decision Tree	0.96	0.95	0.95	0.95

Performance metrics were calculated with the four machine learning algorithm methods used in this study. The success performances of the machine learning algorithms obtained are given in Table 5. According to the model evaluation, F1-Score values were obtained as 0.65 for multinomial naive Bayes, 0.95 for the decision tree, 0.97 for support vector machines and 0.96 for logistic regression classifiers respectively. The performance score of the support vector machine, decision tree and logistic regression is similar.

The figure 2 shows the area under the curve (AUC) of the receiver operating characteristic (ROC) curve analysis of the 4 classification methods. The area under the ROC curve displays similar results which is presented in Table 2. The largest ROC curve area is obtained using support vector machines.

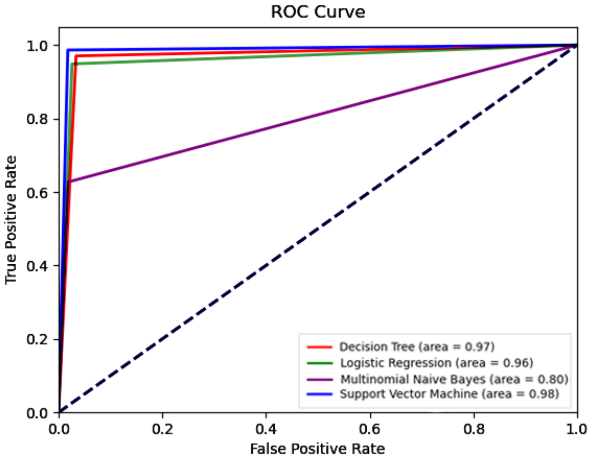

Figure 2. ROC Curves of Four Classification Methods

When the study is compared with similar studies, it is observed that the results obtained are consistent with the literature. According to the obtained results in the study, the algorithm with the best success was obtained by using support vector machines with a 0.97 F1-Score value. In the sentiment analysis study conducted by İlhan et al. [6], the highest success performance was obtained with support vector machines. Likewise, another study [20] achieved the highest success performance by using the support vectors algorithm in their sentiment analysis study.

In addition, as a result of the analysis, the multinomial naive Bayes classifier achieved a lower result than other machine learning algorithms with a successful performance of 0.65 F1-Score. Kaynar [21] and Çoban [22] showed in their studies that the multinomial naive Bayes classifier works with lower performance than other classification algorithms. Considering the obtained results, it was observed that the success rate achieved by the multinomial naive Bayes classifier is approximate with our study.

4. Conclusion

Sentiment analysis is basically a text analysis and aims to determine the class that the given text wants to express emotions as positive, negative and neutral. It is a popular method today and can be used for research in many areas. Generally, it is used to analyze the opinions of the companies about the service they provide or by getting feedback from

the customers who use the products of the brands. In addition, sentiment analysis aims to find out what people think about a specific issue.
In this study, it is aimed to classify people's thoughts about COVID-19 as positive, negative, neither positive nor negative (neutral) by taking the tweets randomly shared by users about COVID-19 on the social media platform Twitter.

The difference of this study from other similar studies in the literature is that we collected our dataset due to ten different keywords. These keywords are determined by the frequency of terms in hashtags and the content of tweets. In addition, tweets of our study are collected at the first phase of pandemia. In the retrieval period, there are lots of different opinions and feelings about COVID-19. The tweets we obtained were taken from the first time of the coronavirus pandemic, so it allowed us to analyze the first feelings and thoughts of people about the pandemic. In pandemia, text mining techniques are applied to tweets to evaluate the opinions of people. The hybrid approaches that utilized TF-IDF weighting, tag clouds, word2vec and various machine learning algorithms are used in literature. The performance results of our study are higher or close to the results of similar studies whose accuracy values are between 0.6 and 0.95 in three label classification [2,4].

Using machine learning algorithms on the dataset obtained from Twitter, it was investigated which algorithm would classify more successfully. As a result, all machine learning algorithms have been compared. It has been observed that the support vector machines algorithm has a high success performance in classification with a success rate of 0.97. The performances results will be improved via obtaining millions of tweets about COVID-19 and classifying them with deep learning algorithms in the future.

References

[1] Barkur, G. and Vibha, G. B. K. (2020). Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India. Asian journal of psychiatry, 51, 102089.

[2] Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R. and Hassanien, A. E. (2020). Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media. Applied Soft Computing, 97, 106754.

[3] Alamoodi, A, et al. (2020). Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. Expert systems with applications, 114155.

[4] Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A. and Choi, G. S. (2021). A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. Plos one, 16(2), e0245909.

[5] Ayan, B., Kuyumcu, B., Ciylan, B. (2019) Detection of Islamophobic Tweets on Twitter Using Sentiment Analysis, Gazi University Journal of Science Part C, 7(2), pp 495-502.

[6] İlhan, N., Sağaltıcı D. (2020) Sentiment Analysis in Twitter, Harran University Journal of Engineering, 5(2), pp. 146-156, doi: 10.46578/humder.772929

[7] Akın, B. ve Şimşek, T. (2018) Adaptive Learning Lexicon Based Sentiment Analysis Proposal, Information Technologies Journal, 11(3), doi: 10.17671/gazibtd.342419

[8] Uslu, A., Tekin, S. ve Aytekin, T. (2019) Sentiment Analyasis In Turkish Film Comments, IEEE 27th Signal Processing and Communications (SIU), doi: 10.1109/SIU.2019.8806355

[9] Neethu, M.S., Rajasree, R. (2013) Sentiment Analysis in Twitter Using Machine Learning Techniques, 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), doi: 10.1109/ICCCNT.2013.6726818

[10] Saha, S., Yadav, J. ve Ranjan, P. (2017) Proposed Approach for Sarcasm Detectionin Twitter, Indian JNournal of Science and Technology, 10(25), pp. 1-8.

[11] Gautam, G., & Yadav, D. (2014). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. 2014 Seventh International Conference on Contemporary Computing (IC3). doi:10.1109/ic3.2014.6897213

[12] Wongkar, M., & Angdresey, A. (2019). Sentiment Analysis Using Naïve Bayes Algorithm of The Data Crawler: Twitter. 2019 Fourth International Conference on Informatics and Computing (ICIC). doi:10.1109/icic47613.2019.8985884

[13] Mandloi, L., & Patel, R. (2020). Twitter Sentiments Analysis Using Machine Learning Methods. 2020 International Conference for Emerging Technology (INCET). doi:10.1109/incet49848.2020.9154183

[14] El Rahman, S. A., AlOtaibi, F. A., & AlShehri, W. A. (2019). Sentiment Analysis of Twitter Data. 2019 International Conference on Computer and Information Sciences (ICCIS). doi:10.1109/iccisci.2019.8716464

[15] Al Shammari, A. S. (2018). Real-time Twitter Sentiment Analysis using 3-way classifier. 2018 21st Saudi Computer Society National Computer Conference (NCC). doi:10.1109/ncg.2018.8593205

[16] Documentation–tweepy, Tweepy. "3.5. 0 documentation." (2020).

[17] Loper, E., & Bird, S. (2002) NLTK: the natural language toolkit. arXiv preprint cs/0205028.

[18] Loria, S., Keen, P., Honnibal, M., Yankovsky, R., Karesh, D., & Dempsey, E. (2014). Textblob: simplified text processing. Secondary TextBlob: simplified text processing, 3.

[19] Pedregosa, Fabian, et al. (2011) Scikit-learn: Machine learning in Python. Journal of machine Learning research 12, 2825-2830.

[20] Çelik, Ö, Osmanoğlu, U, Çanakçı, B. (2020). Sentiment Analysis from Social Media Comments, Mühendislik Bilimleri ve Tasarım Dergisi, 8 (2), 366-374. DOI: 10.21923/jesd.546224

[21] Kaynar, O, Görmez, Y., Yıldız, M. ve Albayrak, A. (2016) Sentiment Analysis with Machine Learning Techniques, Processing Symposium (IDAP'16), pp. 17-18.

[22] Çoban, Ö., Ozyer, B. ve Ozyer, G. (2015) Sentiment Analysis for Turkish Twitter Feeds, 3th Signal Processing and Communications Applications Conference (SIU), pp. 2388-2391.