Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Machine Learning-based Voting Classifier for Improving Sentiment Analysis on Twitter Data

Huatao Li

School of Information Science and Technology, ShanghaiTech University, ShangHai, 201203, China

liht1@shanghaitech.edu.cn

Abstract. As the number of individuals sharing their thoughts on Twitter continues to grow, comprehending the underlying sentiment behind these tweets becomes increasingly crucial for researchers. To identify the optimal model capable of accurately distinguishing tweet sentiment, the author uses a dataset published in 2022, containing tweet texts annotated with corresponding sentiments. Six basic machine learning classification methods are used for model training: Logistic Regression, Naïve Bayes Classifier, Support Vector Classifier, Decision Tree Classifier, Random Forest Classifier, and K-Nearest Neighbors Classifier. Subsequently, the author assesses the trained models. Through the validation, the author finds that the Logistic Regression, Support Vector Classifier, and Random Forest Classifier perform the highest accuracy and F1-score, and the differences between these three models are small. To improve the model, the author votes the best three models together to build a new model. This model's accuracy and F1-score are better than all the basic models, and the accuracy and F1-score have all reached 71.6%. The research shows the differences between each model and the best model when distinguishing between positive tweets, neutral tweets, and negative tweets.

Keywords: Machine learning; sentiment analysis; voting classifier.

1. Introduction

As technology advances and the world becomes increasingly interconnected, the internet attracts a growing number of individuals who desire to express their thoughts online. Some individuals utilize comments to chronicle their lives, showcasing the captivating aspects of their daily existence. Conversely, others employ this platform to vent their grievances and express dissatisfaction. Through the comments they write on the internet, the government can learn about the public's opinions and attitudes towards various topics, products, or events. Businesses and organizations can make informed decisions by knowing how the public thinks about their products, services, or marketing strategies. So, it is important to do sentimental analysis on these texts to know how people think.

The recent research has mostly focused on the results of the model. One of the hot topics these years is COVID-19. Boon-Itt and Skunkan analyze people's opinions about COVID-19's spread trend and divide the stage of it into three different parts [1]. Valerio La Gatta and his team investigated how the virus COVID-19 influences Italy and how the events of the COVID-19 outbreak influence Italian people's daily lives [2], while Yuxing Qi and his team kept their eyes on England [3]. The online class because of the virus is also interesting. Mostafa L. used the same method to research the students' attitudes about this [4]. There is also some research on other topics. Al Amrani Y and his team analyzed people's feelings toward the product through Amazon's product reviews [5]. An analysis of the sentiment of tweets is also a good idea. Adwan OY and his team use different ways of sentimental analysis on Twitter [6], Jarun K and his partner use multilingual Twitter data to do the sentiment analysis [7], and Zahoor S and Rohilla R use an unsupervised method to do the sentimental analysis of tweets [8]. Most of the researchers use the model as a tool to do sentimental analysis for tweets, then research people's attitudes toward different problems, while some of the others pay more attention to the sentimental analysis itself [9-11].

The author thought the research to show the difference between different models when distinguishing the sentiment for the dataset was also important. The selection of a suitable model impacts the project's effectiveness in understanding and categorizing the sentiments expressed in tweets. Different models have varying capabilities for capturing the nuances of language, contextual understanding, and identifying sentiment indicators within the limited text of tweets.

In this research, the author used six different machine learning methods to do the sentimental analysis for the texts of a tweet: Logistic Regression, Naïve Bayes Classifier, Support Vector Classifier, Decision Tree Classifier, Random Forest Classifier, and K-Nearest Neighbors Classifier, and do the validation for each of them to show the differences between them. The dataset was release by M Yasser H, 2022, on Kaggle.

2. Method

2.1. Sentimental Analysis Approach

Tweet sentimental analysis is a process that uses the features of the tweet's text to distinguish if the tweet is negative, positive, or neutral. The author uses the machine learning method to do the sentimental analysis, which can extract feature vectors and build a classifier. The main steps the author takes are: data exploration, data cleaning and preprocessing, vectorizing the texts, using different machine learning methods to train the data, analyzing the results, and finally, making some improvements [6].

2.2. Dataset

The dataset contains 27481 rows and 2 columns. The text column includes 27480 non-null entries, and the sentiment column contains 27,481 non-null entries. The distribution of sentiment is 28.31% for negative, 31.23% for positive, and 40.46% for neutral. The distribution of the three sets of criteria is basically balanced.

Fig. 1, 2, 3 show the word clouds of different sentiments. The negative word cloud is scattered with words such as "miss," "sad," and "sorry." The negative word cloud is scattered with words such as "go," "day," and "use," while the positive word cloud is scattered with words such as "thank," "love,"

Fig. 1 Word cloud of negative tweets (Figure Credit: Original)

Fig. 2 Word cloud of positive tweets (Figure Credit: Original)

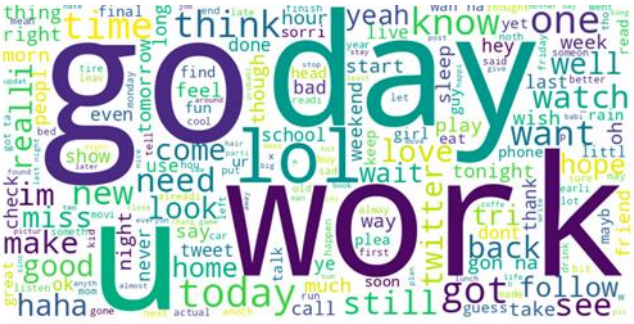

Fig. 3 Word cloud of neutral tweets (Figure Credit: Original)

2.3. Data Preprocessing

Before using the machine learning method to do the sentimental analysis, the preprocessing of the data is indispensable. To let the classifiers understand the texts, the author converted all uppercase letters to lowercase, removed Uniform Resource Locators (URLs), special characters, numbers, and stopped words that were useless for sentimental analysis. After basic data cleaning, the author tokenizes the data to transform the unstructured text data into a structured format, facilitating various downstream tasks and analyses. Part-Of-Speech (POS) tagging is important. It is the process of assigning a specific part of speech to each word in a given text, which can help with feature selection and extraction [12]. Stemming and lemmatization are also necessary. They can use word roots to replace the words to reduce the dimensionality and complexity of the texts. After preprocessing, the datasets can be well utilized and modeled by classifiers.

2.4. Machine Learning Methods

The author vectorized the tweet text as X and labeled the sentiment as Y. To show the relationship between the three sentiments, the author labeled negative as -1, neutral as 0, and positive as 1. Then divide the data into a training set and a validation set randomly, with a ratio of eight to two. Then the author uses different machine learning methods to train the models and predict the sentiment labels. The classification models the author used are as follows:

2.4.1. Logistic Regression

Although there is a "regression" in its name, Logistic Regression is used for classification tasks, especially binary classification tasks, rather than regression. Its principle is using logistic function to model the features' linear combination, producing probability, an output between 0 and 1. By estimating the probability that a given input belongs to a specific category, it can model the label and different features' relationship.

2.4.2. Naïve Bayes Classifier

As its name, Bayesian Classifier using Bayesian theorem to do probability classification. It uses data features' prior probability and likelihood to calculate the probability of data to do the classification. When using a Bayesian classifier, the user assumes that the selected specific feature is independent of other features in terms of existence.

2.4.3. Support Vector Classifier

Support Vector Classifier (SVC) works by finding the optimal hyperplane that best separates different classes in the feature space. In classification problems, the goal of SVC is to find a decision boundary that can separate data points of different categories. Using different kernel function, SVC can deal with different problems, either linear or non-linear.

2.4.4. Decision Tree Classifier

Using eigenvalues, decision tree classifiers can recursively divide the feature space into multiple regions to make decisions. In the decision tree, internal nodes show the judgment of a feature attribute,

branch shows the judgment result, and leaf node shows the output of a category. Decision trees are known for their interpretability and ability to handle both numerical and categorical data.

2.4.5. Random Forest Classifier

It consists of multiple decision trees, each trained independently, and uses random feature selection during the growth process to increase the diversity of the model. When making predictions, random forests determine the final classification result by voting or taking the average of the prediction results of all decision trees. Each tree is randomly trained to avoid homogenization of the generated model.

2.4.6. K-Nearest Neighbors Classifier

K-Nearest Neighbors (KNN) Classifier is a non-parametric, instance-based learning algorithm used for classification and regression tasks. It makes predictions based on the majority class of its K nearest neighbors in the feature space. KNN is simple to understand and implement, making it suitable for various applications. It is particularly effective in handling multi-class classification and non-linear data.

3. Result

This paper uses accuracy, precision, Racall and F1-score as validation metrics. Also, the author draws the Receiver Operating Characteristic curve (ROC) curve for each model used in this research. Mostly, the author only focuses on the accuracy and the ROC curves. The overall comparison is demonstrated in Table 1.

Table 1. Results comparison of different models.

	Accuracy	Precision	Recall	F1-score
Logistic Regression	69.9	70.8	69.9	69.9
Naïve Bayes Classifier	63.9	66.3	63.9	63.7
SVC	70.5	72.3	70.5	70.4
Decision Tree Classifier	65.6	65.5	65.6	65.5
Random Forest Classifier	71.1	71.4	71.1	71.1
K-Nearest Neighbors Classifier	48.8	55.7	48.8	44.3
Vote Classifier	71.7	72.2	71.7	71.7

3.1. Result of Logistic Regression

Logistic Regression is simple, efficient, and interpretable, making it suitable for binary classification tasks. Additionally, it provides probabilities for outcomes, aiding decision-making. However, it has limitations, such as the assumption of linearity, and struggles with complex relationships. It's also prone to overfitting when dealing with high-dimensional data. The performance is shown in Fig. 4. Its accuracy is 69.9%, precision is 70.8%, and recall is 69.9%. Precision and recall are very close, which typically indicates that the model has similar performance in predicting positive and negative instances. The ROC curves show the model performs best when sentiment is positive, then negative, and worst when neutral.

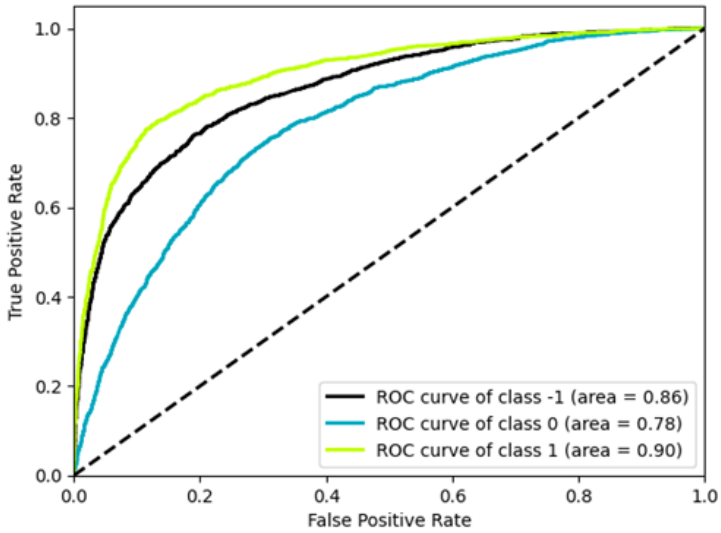

Fig. 4 ROC curves of Logistic Regression model (Figure Credit: Original).

3.2. Result of Naïve Bayes Classifier

Based on Bayes' theorem, Bayesian classifier offers advantages such as handling small data sets, providing probabilistic predictions, and incorporating prior knowledge. It's robust to irrelevant features and performs well in multi-class prediction. However, the model assumes that the features are independent; this situation can never appear in the real world. It's sensitive to the quality of the prior probabilities and may require large amounts of training data for accurate estimation. The ROC performance is shown in Fig. 5.

As a result, when using Naïve Bayes Classifier to train the tweets' sentiment, it performs not so well. The accuracy is only 63.9%. However, the training time of the model is very short. And through the ROC curves, it could be observed that the model performs well both positively and negatively.

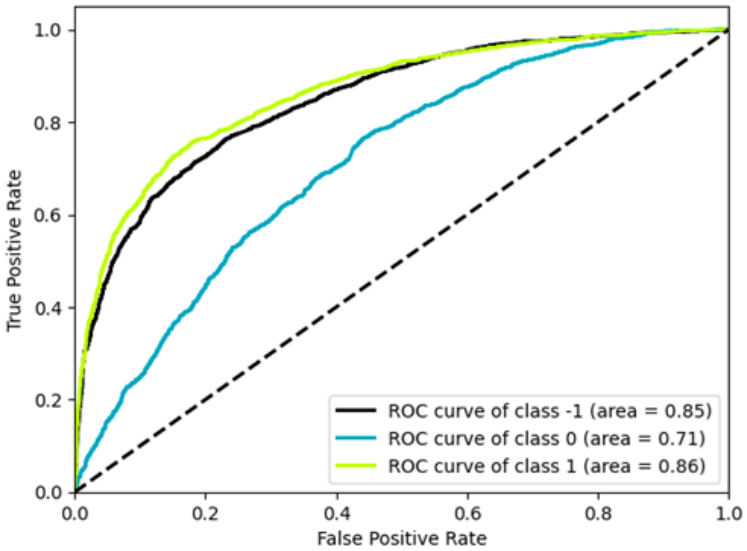

Fig. 5 ROC curves of Naïve Bayes Classifier (Figure Credit: Original).

3.3. Result of Support Vector Classifier

SVC is good at dealing with complex relationships between the features, training high-dimensional data, and providing robust classification. However, it may struggle with large datasets and requires careful selection of kernel functions. Its training time can be significant for extensive datasets, and it's sensitive to the choice of hyperparameters. Its result is demonstrated in Fig. 6.

So, it takes a long time to train the SVC model. But it performs well on this dataset, which has high-dimensional data and complex relationships between words. The ROC curves are almost the same as Logistic Regression, and the accuracy is a little better, at 70.5%.

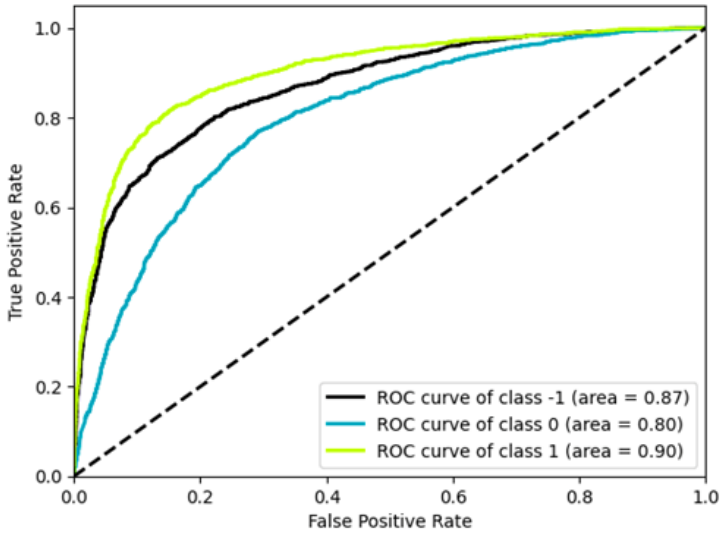

Fig. 6 ROC curves of Support Vector Classifier (Figure Credit: Original).

3.4. Result of Decision Tree Classifier

Decision tree classifier offers interpretability, ease of understanding, and the ability to handle both numerical and categorical data. It's susceptible to overfitting when dealing with complex relationships and noisy data.

Although the accuracy of this model is not bad, which is 65.6%, and the time of training is not long, the ROC curves in Fig. 7 of this model are fold lines, which means the area of each line is not as large as other models. So, Decision tree classifiers are not suitable for complex projects such as sentiment analysis, like this research.

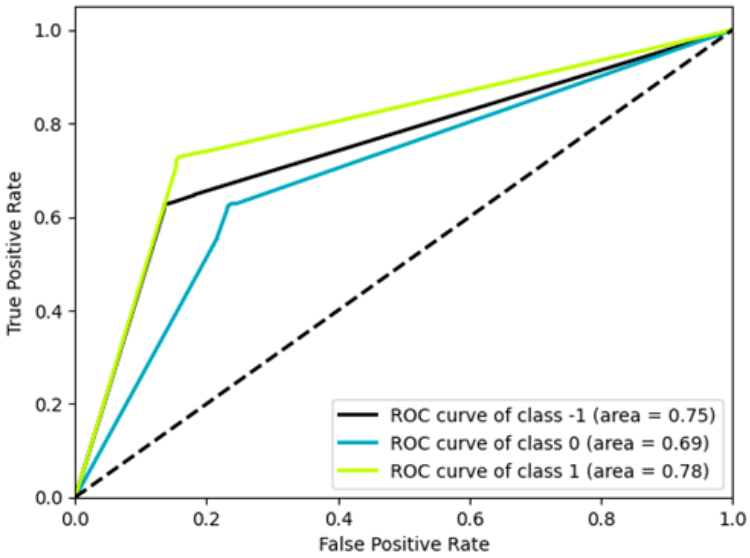

Fig. 7 ROC curves of Decision Tree Classifier (Figure Credit: Original).

3.5. Result of Random Forest Classifier

Random Forest classifier offers high accuracy, robustness to overfitting, and can deal with high-dimensional data perfectly. This classifier can get feature importance from complex relationships. However, it may be computationally expensive for large datasets and complex models. Random

Forest is less interpretable compared to individual decision trees, and it may not perform well on noisy data.

In this research, sentimental analysis with high-dimensional data is where this model performs the best as shown in Fig.8. The accuracy is the largest in these basic models, and the difference between ROC curves is not big, which means the model fits all the sentiments well.

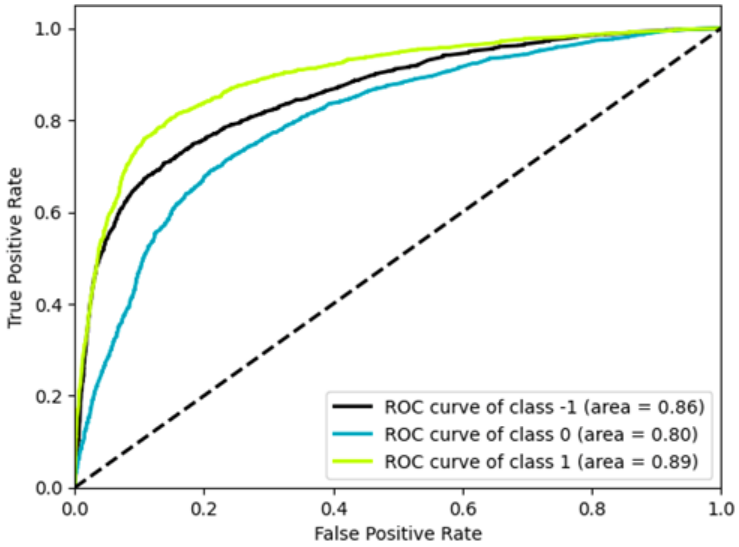

Fig. 8 ROC curves of Random Forest Classifier (Figure Credit: Original).

3.6. Result of K-Nearest Neighbors Classifier

KNN classifier is known for its simplicity, ease of implementation, and effectiveness in handling multi-class classification. It's robust to noisy data and adaptable to new training samples. However, it may struggle with high-dimensional data and be computationally expensive for large datasets.

So, this model is very unsuitable for this process. It can get the same result from the validation metrics and ROC curves. It has low accuracy ROC curves as shown in Fig. 9.

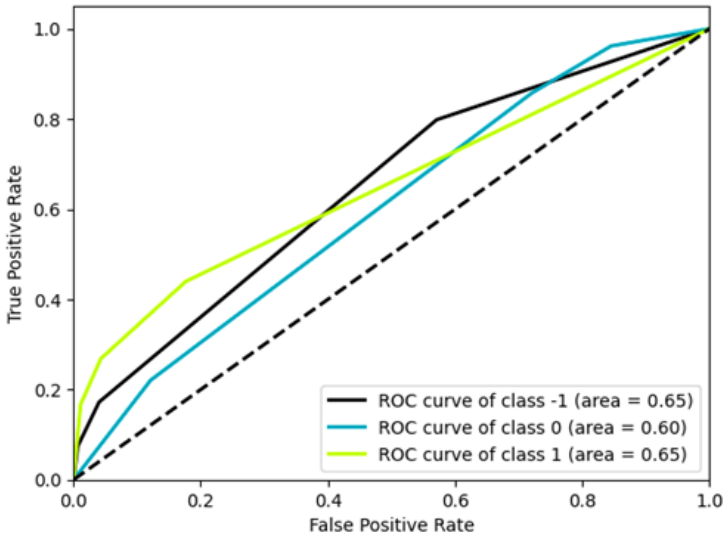

Fig. 9 ROC curves of K-Nearest Neighbors Classifier (Figure Credit: Original).

3.7. Result of Vote Classifier

Through the results of the models above (except the KNN Classifier), the author finds that the accuracy of the models is around 60% to 70%, and the values of precision, recall, and F1-score are very close to their models' accuracy. It indicates that the model is performing consistently across

different evaluation metrics. This shows the model's performance in both positive and negative instances. The closeness of these values implies that the model is making accurate predictions and achieving a good balance between precision and recall.

As displayed in Fig. 10, the ROC curves of these models indicate that they generally have good training effects on positive sentiment, followed by negative sentiment, and have the worst effect on neutral sentiment.

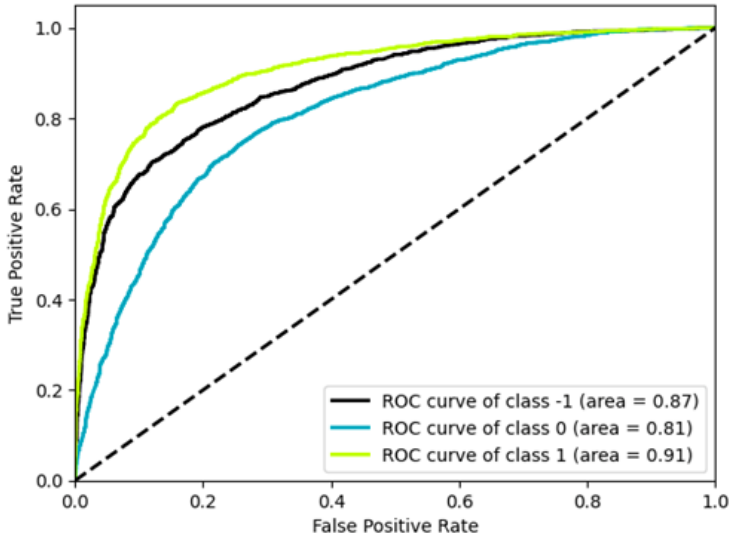

Fig. 10 ROC curves of Vote Classifier (Figure Credit: Original).

The author found that although these models have different principles and methods, there are certain commonalities in the training results. This has led the author to consider using a Classifier to blend the most representative and best-performing models in an attempt to obtain a better model.

The author chooses Logistic Regression, Support Vector Classifier, and Random Forest Classifier to max together to get a better model. The accuracy has improved to 71.7%, and the area of the ROC curves is the largest. The performance is truly improved.

4. Conclusion

In conclusion, the author uses several machine learning methods to do sentimental analysis of tweets and compare the differences between models. Then the author uses a Classifier to max out the best three models and get a model that is better than any other model the author used in this research. The ROC-curves of all the models show an interesting thing: the models' performance on three different sentiments has a common pattern: the performance on positive sentiment is better than negative sentiment, and the performance on negative sentiment is better than neutral sentiment. The author thinks that the models' performances on neutral sentiment are the worst because the number of neutral sentiments is the largest, as the author shows in the data exploration part. It is much bigger than positives and negatives, so there is more noise interference in learning, which leads to poorer learning performance.

However, the positive sentiment is bigger than the negative sentiment, and while the models' performance on positive sentiment is better than that of negative sentiment, the author thinks this is because the data quality of negative sentiment is close to that of positive sentiment, so more data used for training can perform better.

By analyzing the performance of different models, it can also help other researchers make appropriate model choices when conducting related research in order to build more comprehensive models and obtain more accurate results. This not only contributes to scientific research but can also be applied in enterprises to better analyze customer ideas or government agencies in order to better serve the people.

In the future, the author will either do the sentimental analysis process in more interesting directions to learn more about people's ideas or use some other machine learning method that is more powerful, like a neural network or LSTM, to increase the accuracy of the classification of the sentiment behind the text.

References

[1] Boon-Itt, Sakun, and Yukolpat Skunkan. Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study. JMIR Public Health and Surveillance, 2020, 6(4): e21978.

[2] La Gatta, Valerio, et al. COVID-19 Sentiment Analysis Based on Tweets. IEEE Intelligent Systems, 2023 38(3): 51-55.

[3] Qi, Yuxing, and Zahratu Shabrina. Sentiment analysis using Twitter data: a comparative application of lexicon-and machine-learning-based approach. Social Network Analysis and Mining, 2023, 13(1): 31.

[4] Mostafa, Lamiaa. Egyptian student sentiment analysis using Word2vec during the coronavirus (Covid-19) pandemic. Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2020, 2021.

[5] Al Amrani, Yassine, Mohamed Lazaar, and Kamal Eddine El Kadiri. Random forest and support vector machine based hybrid approach to sentiment analysis. Procedia Computer Science, 2018, 127: 511-520.

[6] Adwan, O. M. A. R., et al. Twitter sentiment analysis approaches: A survey." International Journal of Emerging Technologies in Learning (iJET), 2020, 15(15): 79-93.

[7] Arun, K., and A. Srinagesh. Multi-lingual Twitter sentiment analysis using machine learning. Int. J. Electr. Comput. Eng, 2020, 10(6): 5992-6000.

[8] Zahoor, Sheresh, and Rajesh Rohilla. Twitter sentiment analysis using lexical or rule based approach: a case study. 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2020.

[9] Xu, Qianwen Ariel, Victor Chang, and Chrisina Jayne. A systematic review of social media-based sentiment analysis: Emerging trends and challenges. Decision Analytics Journal, 2022, 3: 100073.

[10] Aqlan, Ameen Abdullah Qaid, B. Manjula, and R. Lakshman Naik. A study of sentiment analysis: concepts, techniques, and challenges. Proceedings of International Conference on Computational Intelligence and Data Engineering: Proceedings of ICCIDE 2018, 2019.

[11] Alessia D'Andrea, et al. Approaches, tools and applications for sentiment analysis implementation. International Journal of Computer Applications, 2015, 125(3): 1-8.

[12] Usop, Eka Surya, R. Rizal Isnanto, and Retno Kusumaningrum. Part of speech features for sentiment classification based on Latent Dirichlet Allocation. 2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE). 2017.