Aspect-Based Analysis of Telkomsel User Sentiment on Twitter Using the Random Forest Classification Method and Glove Feature Expansion

Aditya Mahendra Zakaria[1)], Erwin Budi Setiawan[*,1)]

[1)] *Informatics, School of Computing, Telkom University*
Jl. Telekomunikasi No. 1, Terusan Buah batu, Bandung, Jawa Barat, Indonesia 40257

How to cite: A. M. Zakaria, and E. B. Setiawan, " Aspect-Based Analysis of Telkomsel User Sentiment on Twitter Using the Random Forest Classification Method and Glove Feature Expansion," *Jurnal Teknologi dan Sistem Komputer*, vol. 10, no. 2, pp. 114-119, 2022. doi: 10.14710/jtsiskom.2022.14558 [Online].

Abstract – In this modern era, people are certainly very easy to access social media, one of which is Twitter. Twitter is usually used by the public in expressing opinions, be it positive, negative or neutral opinions or can be interpreted as sentiment. The purpose of this study is to analyze the sentiments of Telkomsel users on Twitter using the classification random forest and Glove feature expansion. This study uses an aspect-based sentiment analysis system, namely the signal aspect and the service aspect. Based on the test, it can be concluded that the Random Forest method can provide relevant and accurate classification results for sentiment analysis with the greatest accuracy of 80.37% in the signal aspect and 80.12% in the service aspect, and the expansion feature is proven to increase the performance value of this study by 13.15% on the signal aspect and 5.37% on the aspect service.

Keywords *– sentiment analysis, classification, twitter, random forest , glove feature expansion*

I. Introduction

Telkomsel is one of the largest providers in Indonesia which has a lot of features that can be used by the Indonesian people, for example, internet services. Telkomsel has a very good reputation as a provider operator, which has been operating very well by serving more than 23.5 million customers throughout Indonesia [1].

Various features provided by Telkomsel make people interested in buying, ranging from internet packages, SMS, telephone, multimedia, to digital payments that can be made using Telkomsel. Behind this very complete feature, Telkomsel certainly has several responses from Telkomsel users themselves regarding the positive, negative, and neutral sides related to the features provided by Telkomsel. These responses or arguments are called sentiments [2]. Many arguments on Twitter regarding the features that Telkomsel offers to customers. There are thousands and even millions of tweets every day, this makes it easier for author to get data on Telkomsel user sentiment.

In this study, the author will conduct a sentiment analysis using the random Forest classification method with the expansion feature using Glove. This algorithm is a classifier with an ensemble method consisting of several decision trees to form a random "forest" [3]. Feature expansion is used in this study with the aim of expanding vocabulary which will later assist in the data learning process [4].

Various studies have been carried out to solve the case of sentiment analysis classification on Twitter using various algorithms. Sentiment analysis classification methods that are quite popular and widely used include Naïve Bayes, Decision Tree, Random Forest [5], Alita et al. [6] used the Random Forest classification method for sentiment analysis to detect sarcasm. Sari et al [7] used binary logistic regression, Naïve Bayes classifier (NBC), and Support Vector Machine (SVM) methods. Nasution et al [8] used K-NN and SVM in twitter sentiment analysis.

However [5]-[8] only focus on the standard algorithm (baseline) which has not been optimized to improve its accuracy. There are many ways that can be used to improve the accuracy of a machine learning algorithm for classification, one of which is using the expansion feature. Therefore, this study proposes a Random Forest algorithm for the classification of sentiment analysis of Telkomsel users on Indonesian Twitter with the Glove feature expansion.

In this study, it is divided into 2 aspects, namely the signal aspect and the service aspect. This study will find out how the performance of random forest classification with feature extraction and feature expansion from each aspect, it is hoped that this research can produce better and more accurate performance scores than previous research.

II. Research Methods

A. System Overview

This research begins with crawling data, labeling data, preprocessing data, feature extraction, split data, feature expansion, and modeling sentiment analysis. Here's an explanation of each stage.

[*)] Corresponding author (Erwin Budi Setiawan)
Email: erwinbudisetiawan@telkomuniversity.ac.id

Submitted: 6 September 2021; Revised: 16 December 2021; Accepted: 25 March 2022; Published: 30 April 2022

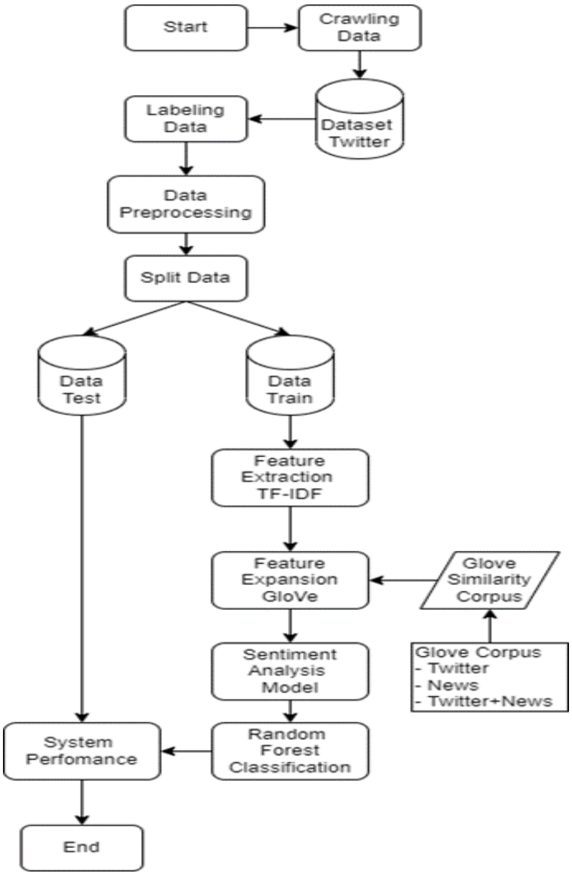

Figure 1. System sentiment analysis using Random Forest Classification and Glove Feature Expansion

B. Data Crawling and Labelling

In this study, datasets were collected from Twitter social media using several keywords related to Telkomsel such as "Telkomsel signal, Telkomsel service, Telkomsel slow, Telkomsel quality, Telkomsel network, Telkomsel down, Telkomsel roaming, etc". The data crawling method using the Twitter API and using the Python Twint library. After crawled data is successfully retrieved, all data is saved in csv file. This data will be used as a training model for sentiment analysis. After crawling, the dataset is labeled with 3 labels, namely positive, negative, and neutral. Can be seen in Table 1 which displays some of the data obtained and has been labeled.

Table 1. Labelling dataset

No	Tweet	Signal	Service
1	Sinyal telkomsel ngajak gelud	-1	0
2	Tumben banget sinyal lancar jaya	1	0
3	Aplikasi Mytelkomsel kenapa error terus	0	-1
4	Aplikasi Mytelkomsel mudahin buat beli kuota	0	1
5	Saya baru proses migrasi	0	0

ke kartu hallo bisa gak
kalo ke grapari saya
batalkan

C. Data Preprocessing

The dataset that has been collected using the crawling method is still in an unstructured state and contains a lot of noise [9]. Therefore, we need a process that can change the form of unstructured data into structured data forms. The preprocessing stage has several processes, namely Data Cleaning, Case Folding, Tokenizing, Stopwords Removing, and Stemming. The following is an explanation regarding each stage of Data Preprocessing:

1. Data Cleaning, remove noise in text such as usernames, hashtags, URL links, numbers, and punctuation.
2. Case Folding, converting capital letters to lowercase.
3. Tokenizing, break the text into tokens or per word contained in the text.
4. Stopword Removing, deleting words in tweets that contain words that are considered not to have an important effect in determining classifications such as conjunctions.
5. Stemming, make it just a root word, removing the suffix and prefix.

D. Data Split

After Preprocessing, the data is continued to split data. Data Split is the process of separating training data and test data [10]. In this study, the proportion of split data used is 80:20 with details of 80% for training data and 20% for test data. The number of training data after being separated is 13,590 and test data is 3,398 data.

E. Extraction Feature TF-IDF

After splitting the data will be taken to the feature extraction stage. Feature extraction is an important factor that can affect the level of accuracy at the classification stage. The selection feature used in this research is TF-IDF. This method works by calculating the weight of each commonly used word [11].

F. Glove Expansion Feature

In this study, the expansion feature used is Glove, the expansion feature used in this study aims to expand the features of a word. The results of the Glove Algorithm will produce an output in the form of a list of similarity words [12]. For example, what can be seen in Table 2 is a list of words that have similarity to the word "internet" which has been sorted according to its rank.

Table 2. Example of Similarity Word From the word ‘Internet’

Rank	Tweet
1	Telkomselgangguan
2	Konek

Submitted: 6 September 2021; Revised: 16 December 2021; Accepted: 25 March 2022; Published: 30 April 2022

3	Akses
4	Pakai
5	Fakta
6	Koneksi
7	Nonton
8	mifi

G. Glove Expansion Feature

Random Forest is a combination of tree predictors. Each tree depends on the value of a random vector whose sample is obtained with a uniform distribution independently for all trees in the forest [13]. Random Forest was introduced by Ho (1995) by combining many trees in the training data to produce a high level of accuracy [14].

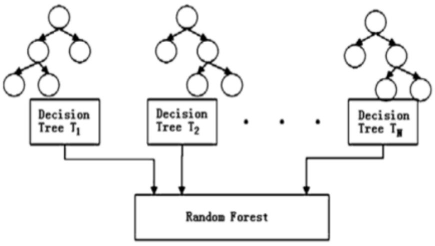

Figure 2. Random Forest visualization

The starting point of the tree is the root node, while the end where the chain ends is called the leaf node. A node represents a particular characteristic, whereas a branch represents a range of values [15]. In the Random Forest partition, the dataset is divided into test and training sets. Each tree will form in-bag data with a subset of the training data and out-of-bag from the remaining parts [16].

H. Hyperparameter Tuning Using Grid Search

Hyperparameter tuning is a critical function necessary for the effective deployment of most machine learning (ML) algorithms. This algorithm has also been used to perform hyperparameter tuning in a case study of tweet emotion classification in Indonesian [17].

In this study. The author conduct experiment with the hyperparameter tuning method to improve performance results. The selection of the right parameters in the model is very important because it aims to improve the results of the performance of the classification itself. The best parameters will produce the best performance values as well. In this study, the hyperparameter tuning method was used to determine the optimizer parameters.

I. Performance Measurement

The results of the classification need to be evaluated to determine the performance of the model that has been made. In this study, the evaluation method used is the confusion matrix. The confusion matrix consists of 4 elements, namely True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN). TP is the proportion of positive labels that are predicted to be correct, FN is the proportion of positive labels that are predicted to be wrong, FP is the proportion of negative labels that are predicted to be wrong, and TN is the proportion of negative labels that are predicted to be correct. There are several performance evaluations that can be calculated using the confusion matrix elements including accuracy, precision, recall, and f1-score [19]. The results of the performance evaluation can be obtained using the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

III. Results and Discussion

A. Data Distribution

Tweet data that has been obtained from the Crawling process amounted to 16,988 Indonesian language tweets containing the opinions of Telkomsel Users in Indonesia. There are 3 labels, namely positive, neutral, and negative, data distribution can be seen in Figure 3.

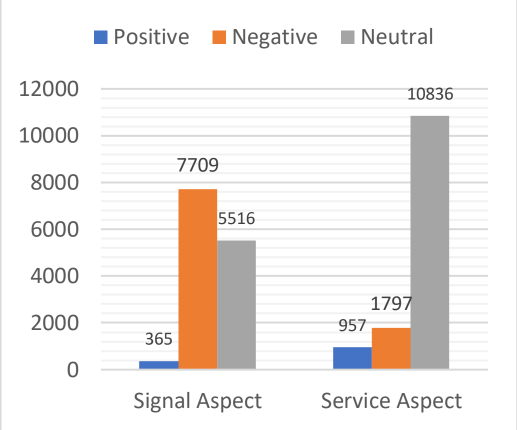

Figure 3. Data distribution amount of each aspect

Then, there is complementary data for making word dictionaries using data taken from several news media such as CNN Indonesia, indonews, Kompas, Tempo, Detik.com, Liputan6, and Republika as much as 142,544 data. The Indonews corpus was used for a combination of experiments in this study to find the best results. There are 3 Corpus Glove created, including Corpus Glove with Tweet dataset, Corpus Glove with News dataset, and Corpus Glove with Tweet+News dataset.

Table 3. Corpus Glove data count

Corpus Glove	Amount
Tweet	16.698
News	142.545
Tweet+News	159.243
Total	**318.486**

Submitted: 6 September 2021; Revised: 16 December 2021; Accepted: 25 March 2022; Published: 30 April 2022

B. Test Results and Analysis

In this study, there are several steps taken before reaching at the evaluation step. The first step is data crawling and data labeling from the tweet data that has been obtained. Second, the previous data will be preprocessed then the data will be splited into training data and test data. After that, the data is extracted using the tf-idf. Then the data is classified using the Random Forest model. The following are some of the scenarios used in this study:

1. Scenario 1: Testing baseline data using different ratios
2. Scenario 2: Testing data oversampling performance using SMOTE
3. Scenario 3: Testing feature expansion performance using Glove
4. Scenario 4: Testing hyperparameter tuning performance.

1. Scenario 1 (Baseline + TF-IDF)

The first scenario is the determination of the baseline or initial data that will be taken to the next test. This test is carried out on tweet data that is already available from the crawling stage in Table 1, the results of the first scenario can be seen in Table 4, this test was chosen to see the best performance value from each proportion of data.

The first step is to divide the data in table 1 into 3 proportions using split data, including a ratio of 80% training data and 20% test data, then a ratio of 90% training data and 10% test data, and a ratio of 70% training data and 30% test data, then the data is trained using the random forest classification method, then compare the results of each proportion to find the best one.

Which can be seen from Table 5, data with a proportion of 80% train data and 20% test data have the best performance values compared to other proportions with an accuracy value of 56.91% and an F1-score value of 24.18% for the Signal Aspect, and an accuracy value of 79.5% and an F1-score value of 29.7% for the Service Aspect.

Table 4. First scenario result

Ratio	Aspect	Accuracy (%)	F1-score (%)
70:30	Signal Service	56.58 79.3	24.09 29.5
80:20	Signal Service	56.91 79.5	24.18 29.7
90.10	Signal Service	56.56	24.5

2. Scenario 2 (Oversampling with SMOTE)

This test is carried out on the best data that has been obtained from the first scenario which can be seen in Table 5. Data with a proportion of 80% for training data and 20% for test data will then be tested using the oversampling method using the SMOTE library. This oversampling aims to make data that is imbalanced in tweet data becomes balanced data [20].

In this scenario, the first step is data with a proportion of 80% train data and 20% test data, oversampling is carried out so that the data becomes balanced as was done in [9], then after the data is balanced, train using the random forest classification method. Can be seen in Table 5, the level of performance increased rapidly, this happened because in the first scenario stage, the processed data was still not comparable (imbalance).

Table 5. Second scenario result

Aspect	Accuracy		F1-Score	
	Before	After	Before	After
Signal	56.91%	80.37% (+23.46)	24.18%	60.17% (+35.99)
Service	79.5%	80.12% (+0.62)	29.7%	45.29% (+15.59)

3. Scenario 3 (Feature Expansion using Glove)

This test is carried out on the best data that has been obtained from the second test which can be seen in Table 6. The data that has been oversampled with SMOTE is continued for feature expansion which aims to expand the vocabulary of a word in the dataset.

In scenario 3, the first step is to train data from every aspect in Table 6. the data trained are tweet data, news data, and tweet+news data which can be seen in Table 4. the data is trained using the glove expansion feature algorithm available in the python library. The data is trained to get the accuracy value of each top-n feature, in this study 4 categories were used, namely top-1 features, top-5 features, top-10 features, and top-20 features.

It can be seen from Table 6 and Table 7, the best performance value in the third scenario is on feature expansion with Corpus Tweet+News with Feature Top-1 with an accuracy value of 83.58% and an F1- score of 67.87% on Signal Aspect and an accuracy value of 77.52% and F1-score 43.81% on Service Aspect.

Table 6. Third scenario result of signal aspect

Feature	Accuracy			
	Before	Corpus Tweet	Corpus News	Corpus Tweet + News
Top 1	80.37%	83.58%	83.22%	84.53%
Top 5	80.37%	83.58%	83.51%	76.63%
Top 10	80.37%	78.55%	81.59%	73.27%
Top 20	80.37%	77.14%	82.59%	77.15%
Feature	F1-Score			
	Before	Corpus Tweet	Corpus News	Corpus Tweet + News
Top 1	60.17%	66.42%	67.34%	67.87%
Top 5	60.17%	66.12%	67.29%	64.05%
Top 10	60.17%	56.42%	66.27%	62.36%
Top 20	60.17%	55.67%	66.12%	61.36%

Submitted: 6 September 2021; Revised: 16 December 2021; Accepted: 25 March 2022; Published: 30 April 2022

Table 7. Third scenario result of service aspect

Feature	Accuracy			
	Before	**Corpus Tweet**	**Corpus News**	**Corpus Tweet + News**
Top 1	80.12%	79.87%	76.5%	77.52%
Top 5	80.12%	76.49%	76.71%	76.52%
Top 10	80.12%	75.79%	75.84%	77.15%
Top 20	80.12%	74.55%	74.81%	75.98%
Feature	**F1-Score**			
	Before	**Corpus Tweet**	**Corpus News**	**Corpus Tweet + News**
Top 1	45.29%	42.01%	44.66%	50.81%
Top 5	45.29%	41.81%	43.91%	48.05%
Top 10	45.29%	42.99%	42.56%	47.81%
Top 20	45.29%	41.34%	41.23%	45.71%

4. Scenario 4 (Hyperparameter Tuning)

This test is carried out on the best data that has been obtained from the previous test scenario, namely the feature expansion data which can be seen in Table 6 and Table 7, the Top-1 feature expansion data in the Tweet+News corpus is the best performance data in the third scenario, then the data will be processed.

At this stage with the Hyperparameter Tuning method. The best parameters are obtained from the best parameter search process using a library from python, namely GridSearchCV [18] which functions to find the best parameter values from a classification. The best parameters can be seen in Table 8.

Table 8. Best parameter

Parameters	Value
bootstrap	false
max_depth	none
max_feature	sqrt
n_estimator	600

In scenario 4 the data is trained by the random forest classification method but the parameters in the random forest algorithm are changed to the parameters in Table 8. the parameter data in Table 8 is obtained through the Grid Search library available in python.

Can be seen in Table 9 where the accuracy performance value increased to 93.52% and the F1-score increased to 73.58% in the signal aspect, and the accuracy value in the service aspect increased 85.49% and the F1-score increased to 66.32% in the service aspects.

Table 9. Fourth scenario result

Aspect	Accuracy		F1-Score	
	Before	**After**	**Before**	**After**
Signal	84.53%	93.52% (+8.99)	67.87%	73.58% (+5.71)
Service	77.52%	85.49% (+7.97)	50.81%	66.32%) (+15.51)

After doing the four test scenarios, it can be concluded that each test can affect the performance of the random forest model created. For the first scenario, from the test results on training data with proportions of 80:20, 90:10, and 70:30, data with a proportion of 80% train data and 20% test data have higher accuracy than the others. The second scenario is testing data by oversampling using SMOTE, in this test the data that is being trained has a high increase in performance value, this is because the data being tested is balanced , of course, it has animpact because the data in the first test is imbalanced same as done on [20]. The third scenario is data testing with Glove feature expansion. the data was tested using 3 corpus, namely corpus tweet, corpus news, and corpus tweet+news, in this test the best data was obtained from corpus tweet+news in the top-1 feature. The performance value at this stage increases same as done on [9] because the data is trained by expanding the features of the corpus that has been created.

IV. Conclusion

Balancing data using a smooth can affect the performance value of each aspect. The implementation of the expansion feature using Glove can improve the performance value for the better. in the signal aspect, the accuracy value increases to 93.52% and the f1-score value increases to 73.58%, in the service aspect the accuracy value increases to 85.49% and the f1-score value increases to 66.32%.

Suggestions for further research can try to use a combination of other feature extraction methods such as Bag of Words (Bow) with other expansion features such as word2vec or FastText, and with other classification methods such as SVM, Naïve Bayes, and others.

References

[1] Emasriani, Felyta, and Reni Rahmadewi. "Analisa Efektifitas Perbaikan Perangkat BTS Telkomsel Karawang dengan iManager u2000 software." CIRCUIT: Jurnal Ilmiah Pendidikan

[2] N. Monarizqa, L. E. Nugroho, and B. S. Hantono. “Penerapan Analisis Sentimen Pada Twitter Berbahasa Indonesia,” Jurnal Penelitian Teknik Elektro dan Teknologi Informasi, Vol. 1. (2014): 151–155.

[3] Amiarrahman, M Rafi, T. Handhika. “Analisis dan Implementasi Algoritma Klasifikasi Random Forest Dalam Pengenalan Bahasa Isyarat Indonesia (BISINDO)”. Prosiding SEMNAS INOTEK (Seminar Nasional Inovasi Teknologi). Vol. 2, No. 1 (2018). pp 083-088.

[4] E. B. Setiawan, D. H. Widyantoro and K. Surendro, "Feature Expansion for Sentiment Analysis in Twitter," 2018 5th International Conference on Electrical Engineering, Computer Science and Informatics. IEEE, 2018 (pp. 509-513), doi:10.1109/EECSI.2018.8752851.

Submitted: 6 September 2021; Revised: 16 December 2021; Accepted: 25 March 2022; Published: 30 April 2022

[5] V.A Fitri*, N Andreswari, A,M, Hasibuan, “Sentiment Analysis of Social Media Twitter with Case of Anti-LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm,” Procedia Computer Science 161. pp. 765-772, 2019.

[6] Alita, Debby, and Auliya Rahman Isnain. "Pendeteksian Sarkasme pada Proses Analisis Sentimen Menggunakan Random Forest Classifier." Jurnal Komputasi Vol. 8, No. 2(2020) : 50-58.

[7] Sari, E.D Nurindah, and Irhamah Irhamah. "Analisis Sentimen Nasabah Pada Layanan Perbankan Menggunakan Metode Regresi Logistik Biner, Naïve Bayes Classifier (NBC), dan Support Vector Machine (SVM)." Jurnal Sains dan Seni ITS Vol. 8, No. 2 (2020): D177-D184.

[8] Nasution, M.R Aziz, and M Hayaty. "Perbandingan Akurasi dan Waktu Proses Algoritma K-NN dan SVM dalam Analisis Sentimen Twitter." Jurnal Informatika Vol. 6, No. 2 (2019): 226-235.

[9] A Febiana, E.B Setiawan, “Hate Speech Detection on Twitter in Indonesia with Feature Expansion Using Glove” Jurnal RESTI Vol. 5, No.6 ISSN 2021 : 1044-1051.

[10] Sreya, Made Dwi Dharma, and Erwin Budi Setiawan. "Penggunaan Metode Glove Untuk Ekspansi Fitur Pada Analisis Sentimen Twitter Dengan Naïve Bayes Dan Support Vector Machine." eProceedings of Engineering Vol. 9, No. 3 (2022).

[11] H. Kumar, B. S. Harish, and H. K. Darshan, “Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method,” Int. J. Interact. Multimed. Artif. Intell. Vol.5, No. 5, 2019: 109. doi: 10.9781/ijimai.2018.12.005.

[12] Nurjannah, Musfiroh, Hamdani, and Indah F Astuti. "Penerapan Algoritma Term Frequency Inverse Document Frequency (TF-IDF) untuk Text Mining." Informatika Mulawarman: Jurnal Ilmiah Ilmu Komputer 8.3 (2016): 110-113.

[13] L. Breiman, “Random forests,” Mach. Learn., p. 1, 2001, doi: 10.1023/A:1010933404324.

[14] T. K. Ho, “Random decision forests,” 1995, doi: 10.1109/ICDAR.1995.598994.

[15] C. R. Sekhar, Minal, and E. Madhu, “Mode Choice Analysis Using Random Forrest Decision Trees,” in Transportation Research Procedia, 2016, p. 6, doi: 10.1016/j.trpro.2016.11.119.

[16] Kuznetsova, Natalia, et al. "Random Forest Visualization." Eindhoven University of Technology (2014).

[17] Raji, I Damilola, et al. "Simple deterministic selection-based genetic algorithm for hyperparameter tuning of machine learning models." Applied Sciences Vol.12, No. 3 (2022): 1186.

[18] Priyadarshini, Ishaani, and C Cotton. "A novel LSTM–CNN–grid search-based deep neural network for sentiment analysis." The Journal of Supercomputing Vol. 77, No. 12 (2021): 13911-13932.

[19] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, “The impact of class imbalance in classification performance metrics based on the binary confusion matrix,” Pattern Recognit., vol. 91, pp. 216–231, 2019, doi: 10.1016/j.patcog.2019.02.023.

[20] Demidova, Liliya, and I Klyueva. "SVM classification: Optimization with the SMOTE algorithm for the class imbalance problem." 2017 6th Mediterranean conference on embedded computing (MECO). IEEE, 2017.

Submitted: 6 September 2021; Revised: 16 December 2021; Accepted: 25 March 2022; Published: 30 April 2022