

Integrating Information Gain methods for Feature Selection in Distance Education Sentiment Analysis during Covid-19

Syamsu Rijal¹, Pandu Adi Cakranegara², Eka Maya S.S. Ciptaningsih³, Putri Hana Pebriana⁴, A. Andiyan⁵, Robbi Rahim⁶

¹ Politeknik Pariwisata Makassar, Makassar, Indonesia

² Universitas Presiden, Indonesia

³ Bina Nusantara University, Management Department, Jakarta, Indonesia

⁴ Universitas Pahlawan Tuanku Tambusai, Indonesia

⁵ Universitas Faletehan, Serang, Indonesia

⁶ Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia

Abstract – Sentiment analysis is a way to automatically understand and process text data to figure out how someone feels about an opinion sentence. If there are too many reviews, it will take a lot of time and they will start to be biased. Sentiment classification tries to solve this problem by putting user reviews into groups based on whether they are positive, negative, or neutral. The dataset comes from Drone Emprit Academic. It is made up of tweets with the words "online learning method" in them, with as many as 4887 data crawled from them. Information Gain and adaboost on the C4.5 (FS+C4.5) method are used in the feature selection method. We use feature options to get rid of bias and improve accuracy. The results of the experiments will be compared to other algorithms like C4.5 and random forest. Based on the results, the accuracy of the two standard decision tree models (C4.5 and random forest) went up from 48.21% and 50.35% to 94.47 %.

The value of how accurate it was went up by 44 percent. The FS+C4.5 model, on the other hand, has an RMSE of 0.204 and a correlation of 0.944. So, adding the feature selection technique to the sentiment analysis of bold learning education can make the C4.5 algorithm even more accurate.

Keywords – Sentiment Analysis, Random Forest, C4.5, Decision Tree, Twitter.

1. Introduction

Sentiment analysis is an automated approach for understanding and processing textual data to extract sentiment information from an opinion sentence [1], [2]. Text mining and natural language processing are often used in conjunction with sentiment analysis (opinion mining) research to evaluate a product [3]. The rapid growth of information technology has led to changes in the way people communicate through the use of social media such as Facebook, Twitter, Instagram, Youtube and Google+ [4], [5], and sentiment analysis is in line with this. This is done in order to facilitate the exchange of information and the expression of ideas and viewpoints on a wide range of topics that are of current interest within the community [6], [7], [8].

In opinion mining, there are two types of sentiment classification techniques: machine learning and lexicon-based techniques [9], [10], [11], [12]. This study focuses on Machine Learning approaches in which classification data mining, which is part of Machine Learning techniques [13], [14], [15] which is used for research [16], [17]. Also, Syamala and Nalini [18] have done a number of studies using a random tree and the adaboost technique to analyze the sentiment of Amazon reviews of products. The results of the experiments show that the proposed model is more effective than common classification

DOI: 10.18421/TEM121-35

<https://doi.org/10.18421/TEM121-35>

Corresponding author: Robbi Rahim,
Sekolah Tinggi Ilmu Manajemen Sukma, Medan,
Indonesia.


Email: usurobbi85@zoho.com

Received: 29 September 2022.

Revised: 08 December 2022.

Accepted: 12 December 2022.

Published: 27 February 2023.

 © 2023 Syamsu Rijal et al; published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License.

The article is published with Open Access at <https://www.temjournal.com/>

models like KNN, Random Forest, and Nave Bayes. The next study Neelakandan & Paulraj [19] looked at how to classify the sentiment of Twitter data using a gradient-boosted method and a decision tree (GBDT). The results of the tests show that GBDT is better than DeepCNN, ANN, Deep Learning NN, and Deep Learning Modifield NN in terms of accuracy, recal, precision, and f-score. Then, research Al-Amrani et al [20] on sentiment analysis using a mix of SVM and decision tree (DTSVM). Based on the results of the experiments, it can be said that the proposed model is more accurate and uses less CPU time than other algorithms like decision trees, SVM, Naive Bayes, PART, and Logistic Regression.

In addition to decision tree, support vector machines (SVM) [21], [22], Naive Bayes [23], [24], [25], neural networks [26], [27], [28], [29], Bayesian networks [30], and maximum entropy [31], [32], [33] are often used in sentiment analysis. You can choose one of these methods based on the problem you want to study [4]. Text mining models have trouble classifying data when there are too many attributes in the model [34], [35]. Another problem is that it's hard to find the best parameters so that the accuracy isn't too high [36], [37]. The new part of this research is the proposed model, which improves the accuracy value by combining the Feature Selection technique with the Decision Tree algorithm. For a higher level of accuracy, the feature selection technique uses Information Gain and adaboost on the C4.5 (FS+C4.5) method on a dataset from Drone Emprit Academic taken from a collection of tweets with the keyword "online learning method."

2. Research Methodology

This research requires a computer with Intel(R) Core (TM) i7-4980HQ 2.80 GHz processor, 16 GB RAM, and Windows 10 Pro operating system. For processing using software assistance from Rapid Miner Studio 9.10. Experimentation and model testing utilize a portion of the existing dataset. The source of the Drone Emprit Academic dataset [38] employs Twitter's API (Application Programming Interface) to capture real-time conversations using the streaming method [38]. The data used in this study were extracted from a corpus of language-related tweets containing the keyword "online learning method." The data crawling phase yielded as many as 4887 data. Data collection occurred between April 1 and May 15, 2022. The dataset's attributes consist of type, mentions, date, link, media, and sentiment.

The study proposes a method for feature selection with information gain parameters, where information gain is used in an integrated manner to improve the classification algorithm's accuracy, and is combined

with adaboost using C4.5 (FS+C4.5). The investigation will yield precision, root mean square error, and correlation. Comparisons will be made between the proposed model and other decision trees, such as C4.5 and random forest. Figure 1 for the proposed model is detailed and condensed below.

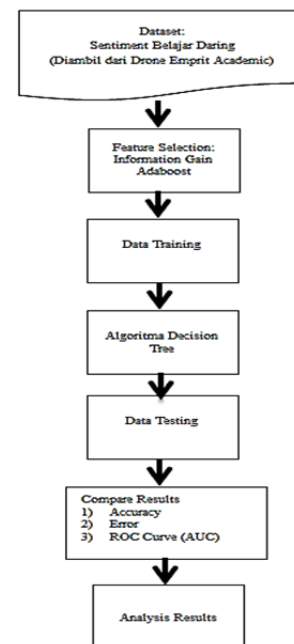


Figure 1. The proposed model

AdaBoost and feature selection are two popular techniques used in sentiment analysis, which is the process of identifying and extracting subjective information from text data.

AdaBoost is an ensemble learning method that combines multiple weak learners to create a strong learner. In sentiment analysis, AdaBoost can be used to improve the performance of the sentiment classification model by combining multiple weak classifiers to create a more accurate and robust model. This is done by repeatedly training the weak classifiers on different subsets of the training data, with the sample weights being adjusted based on the performance of each classifier. The final model is then a weighted combination of the weak classifiers, with the weights being determined by the performance of each classifier.

Pseudo code for AdaBoost in sentiment analysis:

1. Initialize the sample weights for each training example to $1/N$, where N is the total number of training examples.
2. For each weak classifier:
 - a. Train the weak classifier on the training data using the current sample weights.
 - b. Calculate the error rate for the weak classifier on the training data.
 - c. Calculate the weight for the weak classifier based on the error rate.

- d. Update the sample weights for the training examples based on the performance of the weak classifier.
- e. Combine the weak classifiers using the calculated weights to create the final model.

Feature selection is a technique used to identify and select the most relevant and informative features from a dataset for a specific task. In sentiment analysis, feature selection can be used to identify the most important words or phrases in a text that are most indicative of the sentiment expressed. This can help to improve the performance of the sentiment classification model by reducing the number of irrelevant or noisy features in the dataset.

Pseudo code for feature selection in sentiment analysis:

1. Preprocess the text data to remove stop words, punctuation, and other irrelevant elements.
2. Create a list of the most frequently occurring words or phrases in the text data.
3. For each word or phrase in the list:
 - a. Calculate the relevance of the word or phrase to the sentiment classification task using a suitable metric (e.g. mutual information, chi-squared test, etc.).
 - b. Select the top N words or phrases with the highest relevance as the final features.

The novelty of using AdaBoost and feature selection in sentiment analysis lies in the improved performance and robustness of the sentiment classification model. By combining multiple weak classifiers and selecting the most relevant features, the model is able to better capture and classify the sentiment expressed in a given text. This can lead to more accurate and reliable sentiment analysis results. Here the process when using AdaBoos and feature selection perform:

1. Preprocess the text data by removing any irrelevant or noisy features, such as stop words, punctuation, and numbers.
2. Split the preprocessed text data into training and testing sets.
3. Use feature selection to identify the most important words or phrases in the training set that are most indicative of the sentiment expressed.
4. Train a sentiment classification model on the training set, using the selected features as input.
5. Use AdaBoost to combine multiple weak classifiers to create a strong learner.
6. Test the performance of the AdaBoosted sentiment classification model on the testing set.
7. Use the model to classify the sentiment of new text data.

3. Results and Discussion

RapidMiner 9.10 is the software used to assess the sentiment of online learning reviews extracted from the Drone Emprit Academic dataset. The comparison between the suggested model and the classification model is given in Figure 2.

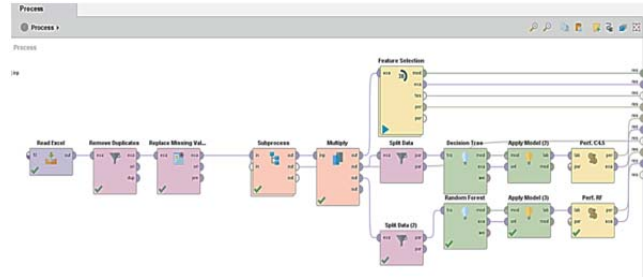


Figure 2. The proposed model with the help of RapidMiner analysis

The dataset from the Emprit Academic Drone that was in (.csv) format was changed to (.xls) format. During the preprocessing stage, you delete rows that are repeated, pick a data set or a single data set that has certain information, and test a text to see if it is true or false. "Remove duplicates," "Select attributes," and "Subprocesses" are the tools that are used. The next step is to choose the features by using the adaboost technique and the C4.5 algorithm together. For the training and testing process, the dataset is split into two parts with a 70:30 split. The "apply model" and "performance C4.5" tools are used to model the training results. During the modeling process, values of accuracy, root mean square error (RMSE), and correlation are found. As shown in Figures 2 and 3, the proposed model will be compared to other decision trees like C4.5 and random forest.

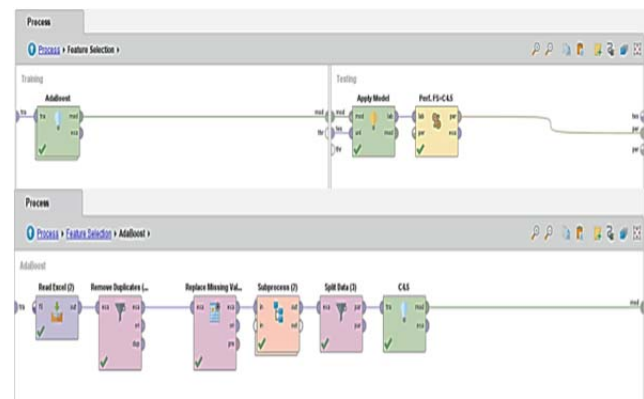


Figure 3. Complete model proposed with RapidMiner

In Table I, you can see the results of the experimental analysis using the two traditional models (C4.5 and Random Forest) and the new model (FS+C4.5).

Table 1. Comparison of accuracy and classification error between models

Parameter	Accuracy	Classification Error
FS+C4.5	94.47%	5.53%
C4.5	48.21%	51.79%
Random Forest	50.35%	49.65%

According to the findings of the confusion matrix test, employing the standard classification model yields an accuracy rating of 48.21% for C4.5 and 50.35% for Random Forest. In contrast, the accuracy rate of the C4.5 algorithm with feature selection (information gain and adaboost approach) is 94.47 percent. The accuracy has increased by 44 percent compared to the previous version. Moreover, the proposed model has an error rate of at least 5.53 percent. The following chart compares the precision and error values of all models, as depicted in Figure 4.

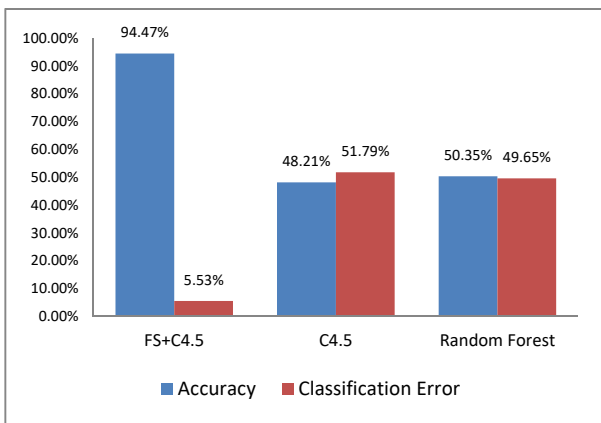


Figure 4. Graph comparison of accuracy and classification error between models

In Table 2, you can see comparison of RMSE and correlation using the two traditional models (C4.5 and Random Forest) and the new model (FS+C4.5).

Table 2. Comparison of RMSE and correlation between models

Parameter	RMSE	Correlation
FS+C4.5	0.204	0.944
C4.5	0.639	0
Random Forest	0.736	0

The correlation coefficient values from highest to lowest are FS+C4.5, C4.5, and Random Forest. The greater the correlation coefficient, the greater the predictive accuracy of the categorization model. 0.944 is the coefficient value in the FS+C4.5 model. On the other side, the greater the accuracy of a classification model, the lower the RMSE number. In these data, the proposed model with an RMSE value of 0.204 has the minimum RMSE value. Figure 5 depicts a graph of the RMSE and correlation for all models.

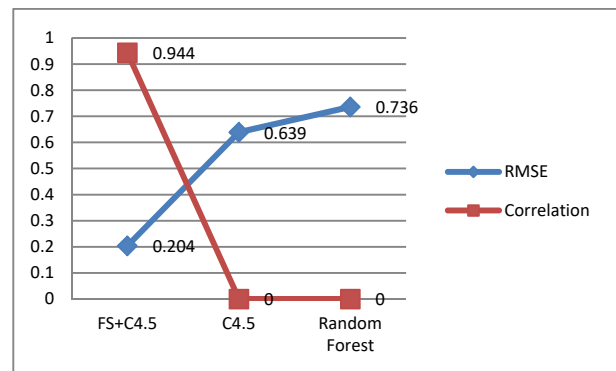
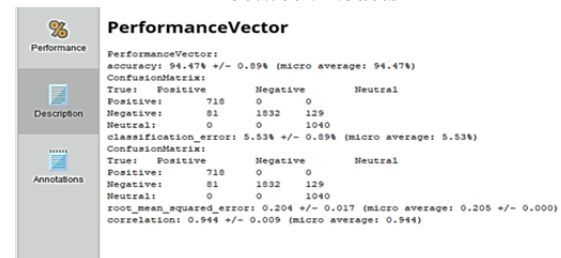
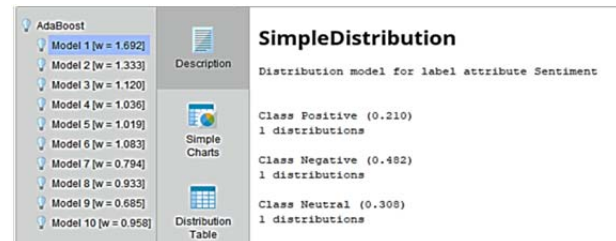


Figure 5. Graph comparison of RMSE and correlation between models



(a)



(b)

Figure 6. Performance recapitulation of the proposed model

The distribution model for the sentiment attribute label is class positive: 1 distributions, class negative: 1 distributions, and class neutral: 1 distributions, with values of 0.210, 0.482, and 0.308, respectively. When compared to the standard version of the decision tree, the C4.5 method with feature selection (information gain and adaboost) gives more accurate results. As Quinlan said, adaboost can be helpful, works better, and is more accurate in its classification.

4. Conclusion

When comparing the C4.5 algorithm and the random forest technique in sentiment analysis, both methods can be effective for classification tasks. The C4.5 algorithm is a decision tree learning method that creates a tree-like model by splitting the data into smaller and smaller subsets based on the values of the features. The random forest technique, on the other hand, is an ensemble learning method that creates a collection of decision trees and combines them to make a prediction.

Both C4.5 and random forest can be effective for sentiment analysis, but the performance of each method will depend on the specific dataset and the parameters used. In general, the random forest technique may be more robust and accurate than the C4.5 algorithm, as it combines multiple decision trees to make a prediction, which can reduce the bias and variance of the model.

In terms of accuracy gain, using the random forest technique instead of the C4.5 algorithm in sentiment analysis can potentially lead to a significant improvement in performance. The exact amount of accuracy gain will depend on the specific dataset and the parameters used in the model, but in general, using random forest can lead to better classification results. The results of the experiments show that if only the standard version of the decision tree is used, accuracy is less than 50%. With FS+C4.5, however, the accuracy goes up by about 40%, to 94.47 percent. It shows that the accuracy value goes up and gets better when C4.5 is used with technique feature selection.

References

- [1]. Nursiyah, S. Y., Erfina, A., & Warman, C. (2021, September). Analisis Sentimen Pembelajaran Daring Pada Masa Pandemi Covid-19 Di Twitter Menggunakan Algoritma Naïve Bayes. In *Seminar Nasional Sistem Informasi dan Manajemen Informatika Universitas Nusa Putra*, 1(1), 117-123.
- [2]. Ningsih, S. R., Manurung, R. T., Bahtiar, A., Firdaus, W., Kusniarti, T., Budiana, N., ... & Sari, V. I. (2018, November). Information System Design E-Assignment in High School to Increase The Effectiveness of Learning. In *Journal of Physics: Conference Series* (Vol. 1114, No. 1, p. 012103). IOP Publishing. doi:10.1088/1742-6596/1114/1/012103
- [3]. Ali, N., Hong, J. E., & Chung, L. (2021). Social network sites and requirements engineering: A systematic literature review. *Journal of Software: Evolution and Process*, 33(4), 1-27. doi:10.1002/smr.2332
- [4]. Samsir, S., Ambiyar, A., Verawardina, U., Edi, F., & Watianthos, R. (2021). Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes. *Jurnal Media Informatika Budidarma*, 5(1), 157-163. doi:10.30865/mib.v5i1.2604
- [5]. Ardiani, L., Sujaini, H., & Tursina, T. (2020). Implementasi Sentiment Analysis Tanggapan Masyarakat Terhadap Pembangunan di Kota Pontianak. *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, 8(2), 183-190. doi:10.26418/justin.v8i2.36776
- [6]. Hemamalini, & Perumal, D. S. (2020). Literature Review on Sentiment Analysis. *International Journal Of Scientific & Technology Research*, 9(4), 2009–2013.
- [7]. Mata, N. P., Mata, N. M., Martins, J. N., Batista, A. C., Rita, J. X. (2021). Sentiment analysis – a literature review. *Academy of Entrepreneurship Journal*, 27(2), 1–10
- [8]. Syafar, F., Gao, J., & Du, J. T. (2013). Applying the international Delphi technique in a study of mobile collaborative maintenance requirements. In *PACIS* (p. 221).
- [9]. Nahar, K. M., Jaradat, A., Atoum, M. S., & Ibrahim, F. (2020). Sentiment analysis and classification of arab jordanian facebook comments for jordanian telecom companies using lexicon-based approach and machine learning. *Jordanian Journal of Computers and Information Technology*, 6(3), 247–262.
- [10]. Naresh Kumar, K. E., & Uma, V. (2021). Intelligent sentiment-based lexicon for context-aware sentiment analysis: optimized neural network for sentiment classification on social media. *The Journal of Supercomputing*, 77(11), 12801-12825. doi:10.1007/s11227-021-03709-4
- [11]. Mamatha, M., Shenoy, R., Thriveni, J., Venugopal, K. R. (2022). Enhanced Sentiment Classification for Dual Sentiment Analysis using BiLSTM and Convolution Neural Network Classifier. *International Journal of Engineering Trends and Technology*, 70(1), 154–163. doi:10.14445/22315381/IJETT-V70I1P217
- [12]. Syafar, F., Husain, H., Ridwansyah, R., Harun, S., & Sokku, S. (2017). Key Data and Information Quality Requirements for Asset Management in Higher Education: A case Study. In *The 30th International Business Information Management Association Conference*, 1670–1677.
- [13]. Bardab, S. N., Ahmed, T. M., Mohammed, T. A. A. (2021). Data mining classification algorithms: An overview. *International Journal of Advanced and Applied Sciences*, 8, (2), 1–5. doi:10.21833/ijaas.2021.02.001
- [14]. Triayudi, A., & Widyarto, W. O. (2021, June). Educational Data Mining Analysis Using Classification Techniques. In *Journal of Physics: Conference Series*, 1933(1), 012061. IOP Publishing. doi:10.1088/1742-6596/1933/1/012061
- [15]. Verma, A. K., Pal, S., & Kumar, S. (2019). Classification of skin disease using ensemble data mining techniques. *Asian Pacific journal of cancer prevention: APJCP*, 20(6), 1887. doi:10.31557/APJCP.2019.20.6.1887
- [16]. Kumar, A., Dabas, V., & Hooda, P. (2020). Text classification algorithms for mining unstructured data: a SWOT analysis. *International Journal of Information Technology*, 12(4), 1159-1169. doi:10.1007/s41870-017-0072-1
- [17]. Yassir, A. H., Mohammed, A. A., Alkhazraji, A. A. J., Hameed, M. E., Talib, M. S., & Ali, M. F. (2020). Sentimental classification analysis of polarity multi-view textual data using data mining techniques. *International Journal of Electrical & Computer Engineering* (2088-8708), 10(5), 5526–5534. doi:10.11591/IJECE.V10I5.PP5526-5534

- [18]. Syamala, M., & Nalini, N. J. (2020). A filter based improved decision tree sentiment classification model for real-time amazon product review data. *International Journal of Intelligent Engineering and Systems*, 13(1), 191-202. doi:10.22266/ijies2020.0229.18
- [19]. Neelakandan, S., & Paulraj, D. (2020). A gradient boosted decision tree-based sentiment classification of twitter data. *International Journal of Wavelets, Multiresolution and Information Processing*, 18(04), 2050027. doi:10.1142/S0219691320500277
- [20]. Al-Amrani, Y. A. S. S. I. N. E., Lazaar, M., & El Kadiri, K. E. (2018). Sentiment Analysis Using Hybrid Method Of Support Vector Machine And Decision Tree. *Journal of Theoretical & Applied Information Technology*, 96(7), 1886–1895.
- [21]. Fitriyaningsih, I., & Basani, Y. (2019). Flood prediction with ensemble machine learning using BP-NN and SVM. *Jurnal Teknologi dan Sistem Komputer*, 7(3), 93-97. doi:10.14710/jtsiskom.7.3.2019.93-97
- [22]. Wable, S., & Laulkar, C. (2013). Fingerprint recognition scheme using assembling invariant moments and SVM. *International Journal of Advanced Research in Computer and Communication Engineering Vol*, 2(10), 3905-3911.
- [23]. Sivakumari, S., Praveena Priyadarsini, R., & Amudha, P. (2009). Accuracy evaluation of C4. 5 and Naive Bayes classifiers using attribute ranking method. *International journal of computational intelligence systems*, 2(1), 60-68.
- [24]. Suseno, H., Wanhari, A., & Masruroh, S. (2020, May). Comparison of C4. 5 and Naïve Bayes Algorithm for Mustahik Classification. In *Proceedings of the 2nd International Colloquium on Interdisciplinary Islamic Studies (ICIIS) in Conjunction with the 3rd International Conference on Quran and Hadith Studies (ICONQUHAS)*, 1–12. doi:10.4108/eai.7-11-2019.2294560
- [25]. Hermanto, H., Kuryanti, S. J., & Khasanah, S. N. (2019). Comparison of Naïve Bayes Algorithm, C4. 5 and Random Forest for Classification in Determining Sentiment for Ojek Online Service. *Sinkron: jurnal dan penelitian teknik informatika*, 3(2), 266-274.
- [26]. Sibyan, H., Suharso, W., Suharto, E., Manuhutu, M. A., & Windarto, A. P. (2021, February). Optimization of Unsupervised Learning in Machine Learning. In *Journal of Physics: Conference Series* (Vol. 1783, No. 1, p. 012034). IOP Publishing.
- [27]. Pratiwi, H., Windarto, A. P., Susliansyah, S., Aria, R. R., Susilowati, S., Rahayu, L. K., ... & Rahadjeng, I. R. (2020, February). Sigmoid activation function in selecting the best model of artificial neural networks. In *Journal of Physics: Conference Series*, 1471(1), 012010. IOP Publishing. doi:10.1088/1742-6596/1471/1/012010
- [28]. Budiharjo, T. S., Windarto, A. P., & Herawan, T. (2018). Predicting tuition fee payment problem using backpropagation neural network model. *Int. J. Adv. Sci. Technol*, 120, 85-96. doi:10.14257/ijast.2018.120.07
- [29]. Budiharjo, T. S., Windarto, A. P., & Herawan, T. (2018). Predicting School Participation in Indonesia using Back-Propagation Algorithm Model. *Int. J. Control Autom*, 11(11), 57-68.
- [30]. Garcia-Diaz, V., Espada, J. P., Crespo, R. G., G-Bustelo, B. C. P., & Lovelle, J. M. C. (2018). An approach to improve the accuracy of probabilistic classifiers for decision support systems in sentiment analysis. *Applied Soft Computing*, 67, 822-833. doi:10.1016/j.asoc.2017.05.038
- [31]. Acosta, M. J., Castillo-Sánchez, G., Garcia-Zapirain, B., De la Torre Diez, I., & Franco-Martín, M. (2021). Sentiment analysis techniques applied to raw-text data from a csq-8 questionnaire about mindfulness in times of COVID-19 to improve strategy generation. *International Journal of Environmental Research and Public Health*, 18(12), 6408. doi:10.3390/ijerph18126408
- [32]. Kabir, M., Kabir, M. M. J., Xu, S., & Badhon, B. (2021). An empirical research on sentiment analysis using machine learning approaches. *International Journal of Computers and Applications*, 43(10), 1011-1019. doi:10.1080/1206212X.2019.1643584
- [33]. Firdaus, W., Eliya, I., & Sodik, A. J. F. (2020). The Importance of Character Education in Higher Education (University) in Building the Quality Students. *Proceedings of the International Conference on Industrial Engineering and Operations Management*, 59, 2602–2606.
- [34]. Saifudin, A., & Wahono, R. S. (2015). Penerapan teknik ensemble untuk menangani ketidakseimbangan kelas pada prediksi cacat software. *IlmuKomputer.com Journal of Software Engineering*, 1(1), 28-37.
- [35]. Wang, Y., & Feng, L. (2020). Improved Adaboost algorithm for classification based on noise confidence degree and weighted feature selection. *IEEE Access*, 8, 153011-153026. doi:10.1109/ACCESS.2020.3017164
- [36]. Qin, C., Liu, S., Lin, S., Li, G., & Hong, J. (2021, October). A Class Imbalance Monitoring Model for Fetal Heart Contractions Based on Gradient Boosting Decision Tree Ensemble Learning. In *International Conference on Data Mining and Big Data* (pp. 217-227). Springer, Singapore.
- [37]. Sonavane, R., & Sonar, P. (2016, December). Classification and segmentation of brain tumor using Adaboost classifier. In *2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC)* (pp. 396-403). IEEE. doi:10.1109/ICGTSPICC.2016.7955334
- [38]. Arianto, B. (2020). Pemanfaatan Aplikasi Drone Emprit Academic dalam Menganalisis Opini Publik di Media Sosial. *Journal of Social Politics and Governance (JSPG)*, 2(2), 177-191.