Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 234 (2024) 349–356

Procedia Computer Science

www.elsevier.com/locate/procedia

Seventh Information Systems International Conference (ISICO 2023)

Comparison of Effectiveness of Logistic Regression, Naive Bayes, and Random Forest Algorithms in Predicting Student Arguments

Tri Wahyuningsih[a,*], Danny Manongga[a], Irwan Sembiring[a], Sutarto Wijono[a]

[a]*Doctoral Program of Computer Science, Universitas Kristen Satya Wacana, Salatiga, Indonesia*

Abstract

Currently, in the process of assessing and giving feedback on students' argumentative writing, educators have to spend a considerable amount of time reading and analyzing each essay individually. This can be a complicated and time-consuming process, especially if the number of students to be assessed is quite large. The problem of this research is to find the most effective algorithm in providing accurate and reliable predictions in the context of evaluation and feedback of students' argumentation. This study compares three algorithms (logistic regression, Naive Bayes, and Random Forest) to predict student argumentation using essays from grades 6-12. Logistic regression performed best with 94.34% accuracy, followed by random forest with 91.98% accuracy, and Naive Bayes with 88.93% accuracy. The study optimized preprocessing and selected algorithms for an automated guidance model. It is the first stage of a three-part study for developing automated guidance models. Data came from Kaggle, and the study aims to improve the accuracy of automated guidance models for student argumentation.

Keywords: Comparison Algorithm; Logistic Regression; Naive Bayes; Random Forest; Student Arguments.

1. Introduction

In an educational context, arguments are an important part of the learning process as they can help students to develop critical thinking and logic skills. Arguments can also help students to understand the information provided and strengthen their beliefs. However, not all students have the ability to make strong and well-structured arguments.

* Corresponding author. Tel.: +6285215405204; fax: +6285215405204.
E-mail address: 982022001@student.uksw.edu

1877-0509
10.1016/j.procs.2024.03.014

Some factors that can affect students' ability to make good arguments include lack of knowledge about the topic being debated, lack of critical thinking skills and logic, and lack of experience in debating.

To address the problem, research was conducted in developing a prediction system that can assist teachers and students in improving students' argument skills. In this prediction system, students' data related to their arguments will be collected and analyzed to identify patterns and characteristics of strong arguments. By studying these patterns and characteristics, the system can predict the likelihood of success of students' arguments and provide recommendations to improve the weaknesses in their arguments. Methods that can be used in developing this prediction system include text analysis and natural language processing. Text analysis is used to identify the structure and content of student arguments, while natural language processing is used to identify patterns and characteristics in human language contained in student arguments.

In the development of increasingly advanced technology, the development of prediction systems to predict the quality of students' arguments will become increasingly important in the context of education. By using the right prediction system, teachers can provide more specific feedback to students to help them improve their critical thinking and logic skills, and strengthen their confidence in making strong and well-structured arguments.

Teachers are one of the professions that have an important role in the learning process. In addition to delivering subject matter to students, teachers should also be able to identify the meaning and quality of students' arguments in their dreams [1,2]. Teachers have difficulty in doing this. There are several reasons why teachers find it difficult to identify the meaning and quality of students' arguments, namely: 1). Students have significant differences in background and goals. For some students, their dreams are related to academic achievements or professional careers, while for other students they are related to sports achievements or other social activities [3–6]. 2). Students have different language abilities. Students who have good language skills find it easier to convey their dreams and arguments in a clear and structured manner. Students who have weak language skills find it difficult to convey their dreams and arguments. 3). Students often have a vague understanding of what they want. For some students, their dreams are only focused on short-term goals, so they do not have a clear vision of what they want to achieve in the long term [7,8].

This research is the first of three stages of research aimed at developing an automated guidance model. At this stage, the main objective is to find the basic or main algorithm that will be used in the development of the model. This research is the first in the series and aims to evaluate the performance of three algorithms used to predict student argumentation automatically. In the next stage, the results of this research will be used as a basis for developing a more complex and accurate automated guidance model. Therefore, this research makes an important contribution to the development of automated guidance models and is an important initial stage in achieving the ultimate goal of this research. This research compares three algorithms namely logistic regression, naive bayes, and random forest. This research tests and proves the role of standard preprocessing in improving the accuracy of each algorithm to be tested. This research will select the algorithm with the best results that will be used as the basic algorithm of the automated guidance model in further research.

2. Literature Review

The development of automated guidance and models for text or argument identification has been an active area of research in recent years. Here are some of the previous expert research results in this field:

Itoo and Singh [9] conducted a study that compared and analyzed the effectiveness of three machine learning algorithms (logistic regression, Naïve Bayes, and KNN) to detect credit card fraud. This study used a dataset containing credit card transactions that included information such as the number of transactions, location, and category of products purchased. The results of this study show that the logistic regression algorithm has the best performance in detecting credit card fraud, followed by the Naïve Bayes and KNN algorithms. However, although the logistic regression algorithm performed the best in this study, the authors emphasize that the selection of the appropriate machine learning algorithm is highly dependent on the characteristics and needs of the dataset used.

Similar results were found in the study of Shah et al. [10] which aims to compare the performance of three different text classification models: Logistic Regression, Random Forest, and K-Nearest Neighbor (KNN). This study used three different datasets to test the performance of these models. The results show that the Logistic Regression model has better accuracy than the Random Forest and KNN models on all three datasets used in this

study. However, the Random Forest model also showed good performance and was significantly better than the KNN model. In addition, this study also shows that the performance of these models can be affected by the size of the dataset and the number of features used in the model. In conclusion, this study shows that the Logistic Regression model can be a good choice for text classification tasks, especially on large datasets with a large number of features. However, the Random Forest model can also be a good alternative and more effective than the KNN model.

Chakrabarty et al. [11] developed a tool called Ampersand to identify and analyze persuasive arguments in online discussions. The researchers used a dataset consisting of 127,086 comments from political discussion forums on Reddit, focusing on discussions related to two controversial political topics: gun control and internet neutrality. Ampersand uses natural language processing techniques to automatically identify argument components such as premises, claims, and refutations. The tool then assigns a score to each component based on its relevance to the discussion and its persuasive power. The researchers evaluated Ampersand's performance using manual annotations of argument components and found that the tool achieved high precision and recall in identifying arguments. They also used the tool to analyze the persuasive strategies used by participants in online discussions and found that participants often used emotional appeals and personal anecdotes to support their arguments. Overall, this study demonstrates the effectiveness of using argument mining tools such as Ampersand to analyze persuasive arguments in online discussions, which can provide insights into how people use language to persuade others and how these discussions can be improved.

Research by Hassani et al. [12] is a literature review that aims to provide an overview of the use of text mining techniques in big data analytics. In this study, the researchers identified and evaluated text mining methods used in big data analytics, such as data preprocessing, document representation techniques, document clustering and classification techniques, and result evaluation. The results of this study show that text mining can provide valuable information and can be used to make decisions in various fields, such as business, healthcare, and cybersecurity. However, researchers also identified some challenges in using text mining in big data analytics, such as complex data processing and long computation time. Overall, this research provides useful insights into the use of text mining in big data analytics and shows the great potential of this technique in providing valuable information to decision makers.

Previous research [13] also shows that Logistic Regression is more suitable for use on less complex data, while Random Forest is more suitable for use on more complex data. However, when compared to interpretability, Naive Bayes is easier to understand compared to Logistic Regression and Random Forest, as Naive Bayes uses simpler probability principles. Overall, the most suitable algorithm to use depends on the complexity of the data and the need for interpretation. If accuracy is the top priority, then Random Forest is the best choice, whereas if the need for interpretation is higher, then Naive Bayes is a more suitable choice. Logistic Regression algorithm has a good ability to predict binary classes (yes or no), but not so good in predicting multi-class classes [7]. Naive Bayes algorithm has a good ability to predict multi-class classes, but not so good in handling features that are highly related to each other. The Random Forest algorithm has a good ability to handle highly related features, but takes longer to train than the Logistic Regression and Naive Bayes algorithms. Random Forest algorithm is proven to have higher accuracy than Logistic Regression and Naive Bayes algorithm in some cases, but it is not always the best choice depending on the data characteristics and intended use.

3. Methodology

3.1. Data Collection

This study uses a dataset taken from Kaggle, namely Feedback Prize - Predicting Effective Arguments: Rate the effectiveness of argumentative writing elements from students grade 6-12. This dataset consists of argumentative essays written by US students in grades 6-12. Each essay was annotated by expert raters to rate the imaginative elements commonly present in argumentative writing. The data consists of two main classes, namely argument type and argument effectiveness. However, in this study, only focus on the argument effectiveness class in the computational process. Table 1 is a sample of the dataset.

Table 1. Sample dataset.

discourse_id	essay_id	discourse_text	discourse_type	discourse_effectiveness
0013cc385424	007ACE74B050	Hi, I'm Isaac, I'm going to be writing about how this face on Mars is a natural landform or if there is life on Mars that made it. The story is about how NASA took a picture of Mars and a face was seen on the planet. NASA doesn't know if the landform was created by life on Mars, or if it is just a natural landform.	Lead	Adequate
9704a709b505	007ACE74B050	On my perspective, I think that the face is a natural landform because I dont think that there is any life on Mars. In these next few paragraphs, I'll be talking about how I think that is a natural landform.	Position	Adequate
c22adee811b6	007ACE74B050	I think that the face is a natural landform because there is no life on Mars that we have discovered yet.	Claim	Adequate
a10d361e54e4	007ACE74B050	If life was on Mars, we would know by now. The reason why I think it is a natural landform is because, nobody lived on Mars in order to create the figure. It says in paragraph 9, "It's not easy to target Cydonia," in which he is saying that its not easy to know if it is a natural landform at this point. In all that they're saying, its probably a natural landform.	Evidence	Adequate
db3e453ec4e2	007ACE74B050	People thought that the face was formed by aliens because they thought that there was life on Mars.	Counterclaim	Adequate

3.2. Data preprocessing

This research uses Python and Jupyter Notebook as tools to perform computations related to the research. Python is a popular and frequently used programming language in the field of data science and software development. In this context, Python was used to implement the algorithms and data analysis required in the research. Jupyter Notebook, as an interactive development environment, makes it easy to write and run Python code and integrate it with explanatory text and visualizations. By using this combination, the research can be conducted efficiently and the computational results can be displayed clearly and easily understood.

Data preprocessing is a data processing process carried out with the aim of making the data ready for analysis [14,15]. This research combines the general preprocessing process with text pre-processing, which consists of several stages, namely case folding, symbol removal, remove duplicates, tokenizing, stop word removal, stemming.

3.3. Evaluation

The algorithms used to develop the Automated Guidance System are Logistic Regression, Naive Bayes, and Random Forest algorithms. First, the training data set and test data set were tested. The training data is used to train the model and the test data is used to evaluate the model's ability to accurately categorize student arguments. Furthermore, the results of the algorithm testing were analyzed using several evaluation matrices, namely accuracy, precision, recall and F1 Score. Accuracy measures how good the model is at classifying student arguments as a whole [16–19], precision measures how good the model is at classifying correct student arguments as correct student arguments, and recall measures how good the model is at finding all correct student arguments.

The results of the analysis are compared to determine which algorithm is the best in developing an Automated Guidance System for classifying student arguments. The algorithm that has the highest evaluation matrix value will be selected as the best algorithm to use in the development of the system. However, it is not always the algorithm with the highest evaluation matrix value that is best to use because sometimes other algorithms that have lower evaluation matrix values can give better results under certain conditions. Therefore, it is important to evaluate all the selected algorithms and compare them carefully before deciding which algorithm to use in system development. The flow of this research is shown in Figure 1.

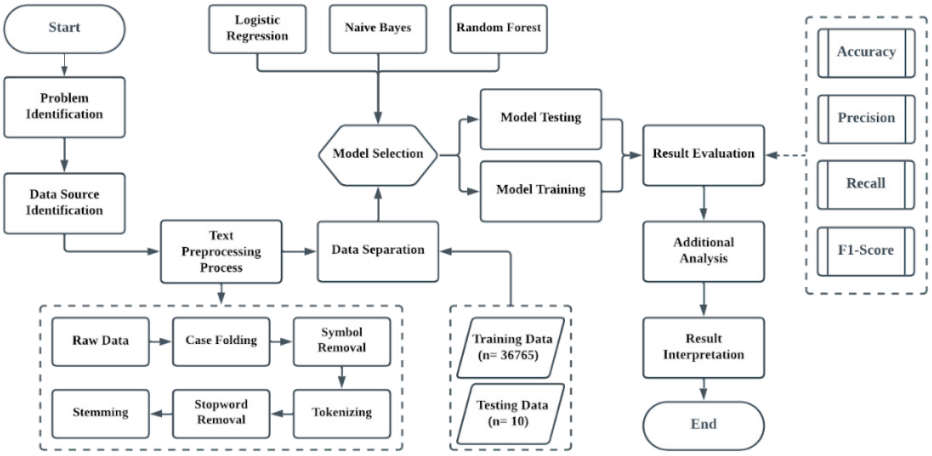

Fig. 1. Flow of the research method.

4. Result and Discussion

4.1. Data Evaluation

The author chose three algorithms, namely logistic regression, naive bayes, and random forest, to compare in this study for certain reasons. First, logistic regression was chosen because it is a commonly used algorithm in classification modeling. It has easy interpretation and is suitable for classifying data with binary target variables. Second, naive bayes was chosen because it is a fast and efficient method in classification modeling. The naive bayes algorithm is suitable for classifying data with statistically independent predictor attributes. Finally, random forest was chosen because it is an ensemble algorithm that is able to combine the results of multiple decision trees to improve prediction accuracy. It can overcome the problem of overfitting and bias-variance tradeoff. By comparing these three algorithms, this research aims to evaluate the effectiveness of each algorithm in the context of automatically classifying student arguments.

In this study, the effectiveness of arguments can be categorized into three namely Effective, Ineffective, and Adequate. Details regarding the composition of the classes that have been identified in the dataset can be seen in Figure 2.

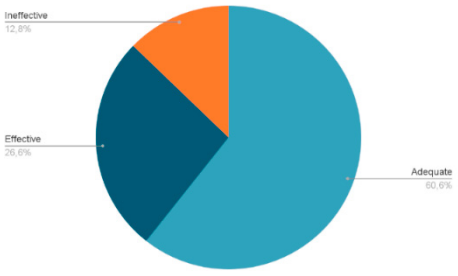

Fig. 2. Class composition.

Figure 4 shows that the most argument effectiveness in the dataset is Adequate (60.6%), followed by Effective (26.6%) and Ineffective (12.8%). These results can be used as a basis for evaluating argument effectiveness in similar texts.

4.2. Data preprocessing

The data is preprocessed by converting to lowercase, removing unnecessary symbols, removing duplicate words, splitting the text into tokens, removing unnecessary words and converting words into root words [20,21]. Table 2 shows the preprocessing simulations used on this research dataset. This process is carried out to change the dataset to suit the needs of the analysis to be carried out.

Table 2. Simulated pre-processing.

	Argument
Raw Data	If life was on Mars, we would know by now. The reason why I think it is a natural landform is because, nobody lived on Mars in order to create the figure. It says in paragraph 9, "It's not easy to target Cydonia," in which he is saying that its not easy to know if it is a natural landform at this point. In all that they're saying, its probably a natural landform.
Case Folding	if life was on mars, we would know by now. the reason why i think it is a natural landform is because, nobody lived on mars in order to create the figure. it says in paragraph 9, "it's not easy to target cydonia," in which he is saying that its not easy to know if it is a natural landform at this point. in all that they're saying, its probably a natural landform.
Symbol Removal	if life was on mars we would know by now the reason why i think it is a natural landform because nobody lives on mars in order to create the figure it says in paragraph 9 its not easy to target cydonia in which he is saying that its not easy to know if it is a natural landform at this point in all that they're saying its probably a natural landform
Remove Duplicate	if life was on mars we would know by now the reason why i think it is a natural landform because nobody lives on mars in order to create the figure says paragraph 9 not easy target cydonia. he is saying that this point. all they are saying, probably
Tokenizing	"if", "life", "was", "on", "mars", "we", "would", "know", "by", "now", "the", "reason", "why", "i", "think", "it", "is", "a", "natural", "landform", "because", "nobody", "lives", "mars", "in", "order", "to", "create", "the", "figure", "says", "paragraph", "9", "not", "easy", "target", "cydonia", "he", "saying", "that", "this", "point", "all", "they", "are", "saying", "probably"
Stopword Removal	"target", "cydonia", "saying", "not", "easy", "point", "probably", "natural", landform"
Stemming	"target", "cydonia", "say", "not", "easy", "point", "probably", "nature", "landform"

4.3. Algorithm Evaluation

Logistic Regression, Naive Bayes, and Random Forest are three algorithms that are often used in data processing. Each of these algorithms has different advantages and disadvantages. In this study, the effectiveness of Logistic Regression, Naive Bayes, and Random Forest algorithms is measured by calculating the accuracy rate of each algorithm. The algorithm that has the highest accuracy rate is chosen as the best algorithm in predicting student argumentation automatically. In addition, this research also evaluates using confusion matrix, precision, recall, and F1 score to compare the effectiveness of each algorithm. The results of accuracy, Precision, Recall and F1 Score can be seen in Table 3.

Table 3. Accuracy, Precision, Recall and F1 Score results.

Algorithm	Accuracy	Precision	Recall	F1 Score
Logistic Regression	94.34%	93.72%	94.96%	94.33%

Naive Bayes	88.93%	88.75%	89.12%	88.94%
Random Forest	91.98%	91.68%	92.31%	91.98%

Table 3 shows the results of accuracy, precision, recall, and F1 Score for each algorithm. To see a clearer comparison between the performance of the algorithms, it is accompanied by Figure 4 which shows the comparison visually through a diagram.

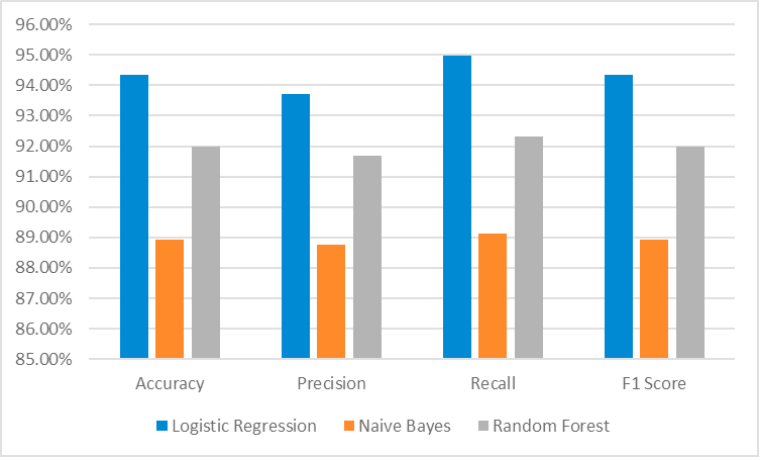

Fig. 3. Result comparison.

Figure 3 shows that the logistic regression algorithm has an accuracy rate of 94.34%, the naive bayes algorithm has an accuracy rate of 88.93%, while the random forest algorithm has an accuracy rate of 91.98%. The curation results show that the logistic regression algorithm has a higher accuracy rate than naive bayes and random forest. The precision result of Logistic Regression is 93.72%, Naive Bayes is 88.75% and Random Forest is 91.68%. This means that of the three algorithms tested, Logistic Regression has the highest precision with a value of 93.72%, followed by Random Forest with a value of 91.68%, and Naive Bayes with a value of 88.75%. The results of recall and FI Score also show that Logistic Regression has better performance.

5. Conclusion

This discovery has potential implications for educators and educational technology developers. Knowing that the Logistic Regression algorithm is superior in classifying student arguments, educators can use automated guidance systems based on this algorithm to provide more accurate and effective feedback to students. This can help improve the learning process and assist students in developing their argumentation skills. In addition, educational technology developers can utilize the results of this research to design and develop better automated guidance systems in terms of classifying students' arguments.

However, this research also has some limitations that need to be considered. One of them is the use of Naive Bayes and Random Forest algorithms as a limited comparison. Further research could involve more algorithms and modeling methods to gain a more comprehensive understanding of the classification of student arguments. In addition, this research may also consider other variables that may affect the modeling of automated guidance systems, such as cultural context or language used. Further research that addresses these limitations may provide further insights into the use of algorithms in the classification of student arguments and the potential development of more sophisticated automated guidance systems in the future.

References

[1] Aziz A, Zaki R, Ullah SE. The Impact of Dynamic Capabilities and Information System on Organizational Effectiveness in Cellular Communication Companies in Jordon: Mediating Role of Organizational Ambidexterity. IJIIS Int J Informatics Inf Syst 2020;3:81–93. https://doi.org/10.47738/ijiis.v3i2.68.

[2] Hulliyah K. Predicting Airline Passenger Satisfaction with Classification Algorithms. IJIIS Int J Informatics Inf Syst 2021;4:82–94. https://doi.org/10.47738/ijiis.v4i1.80.

[3] Umami I. Comparing Epsilon Greedy and Thompson Sampling model for Multi-Armed Bandit algorithm on Marketing Dataset. J Appl Data Sci 2021;2:14–26. https://doi.org/10.47738/jads.v2i2.28.

[4] Riyanto R. Training Autonomous Vehicles in Carla model using Augmented Random Search Algorithm. J Appl Data Sci 2021;2:27–35. https://doi.org/10.47738/jads.v2i2.29.

[5] Nguyen CTD, Trang TTN. Analysis of Internal Factors Effecting Bank Profitability: Evidence From Listed Bank on Vietnam Stock Market. IJIIS Int J Informatics Inf Syst 2021;4:138–48.

[6] Widiyanto A, Prabowo NA, Ircham M, Amarullah N, Soni A. The Effect of E-Learning as One of the Information Technology-Based Learning Media on Student Learning Motivation. IJIIS Int J Informatics Inf Syst 2021;4:123–9.

[7] Akmal. Predicting Dropout on E-learning Using Machine Learning. J Appl Data Sci 2020;1:29–34. https://doi.org/10.47738/jads.v1i1.9.

[8] Kokasih MF, Paramita AS. Property Rental Price Prediction Using the Extreme Gradient Boosting Algorithm. IJIIS Int J Informatics Inf Syst 2020;3:54–9. https://doi.org/10.47738/ijiis.v3i2.65.

[9] Itoo F, Singh S. Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. Int J Inf Technol 2021;13:1503–11.

[10] Shah K, Patel H, Sanghvi D, Shah M. A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augment Hum Res 2020;5:1–16.

[11] Chakrabarty T, Hidey C, Muresan S, McKeown K, Hwang A. Ampersand: Argument mining for persuasive online discussions. ArXiv Prepr ArXiv200414677 2020.

[12] Hassani H, Beneki C, Unger S, Mazinani MT, Yeganegi MR. Text mining in big data analytics. Big Data Cogn Comput 2020;4:1.

[13] Raza GM, Butt ZS, Latif S, Wahid A. Sentiment Analysis on COVID Tweets: An Experimental Analysis on the Impact of Count Vectorizer and TF-IDF on Sentiment Predictions using Deep Learning Models. 2021 Int Conf Digit Futur Transform Technol ICoDT2 2021 2021. https://doi.org/10.1109/ICoDT252288.2021.9441508.

[14] Tomura N. Construction of The E-Government Case Study of Japan and Estonia. Int J Appl Inf Manag 2021;1:145–51. https://doi.org/10.47738/ijaim.v1i3.16.

[15] Hitoshi H. The Effectiveness of The Body of Knowledge Process in The Startup Analysis of Efficiency by Applying Startup Management Body of Knowledge (SUBOK) Guide. Int J Appl Inf Manag 2021;1:28–49. https://doi.org/10.47738/ijaim.v1i2.11.

[16] Huyen Tran TT. Factors Affecting The Collaborative Relationships in Tourism Supply Chain. Int J Appl Inf Manag 2021;1:152–64. https://doi.org/10.47738/ijaim.v1i4.17.

[17] Alomari MS. The Legal System for the Conversion of Commercial Companies in the Light of the Rules of the Saudi Corporate System. Int J Appl Inf Manag 2022;2:106–11. https://doi.org/10.47738/ijaim.v2i4.43.

[18] Wahyuningsih T. Text Mining an Automatic Short Answer Grading (ASAG), Comparison of Three Methods of Cosine Similarity, Jaccard Similarity and Dice's Coefficient. J Appl Data Sci 2021;2:45–54. https://doi.org/10.47738/jads.v2i2.31.

[19] Al-Shoteri A. The Role of Methods and Applications of Artificial Intelligence Tools in the Field of Medicine to Diagnose and Discover Various Diseases. J Appl Data Sci 2022;3:01–14. https://doi.org/10.47738/jads.v3i1.48.

[20] Chang C, Sung C, Lin T. Density-Aware Data Augmentation for Unsupervised Domain Adaptation on Single Image Dehazing (DAMix). Cvpr 2021.

[21] Hananto A. COVID-19 Vaccination: A Retrospective Observation and Sentiment Analysis of the Twitter Social Media Platform in Indonesia. IJIIS Int J Informatics Inf Syst 2022;5:56–69. https://doi.org/10.47738/ijiis.v5i1.126.