Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 407−413
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4079

Prediction Retweet Using User-Based and Content-Based with Artificial Neural Network-Harmony Search

Rizky Ahmad Saputra, Jondri*, Kemas Muslim Lhaksmana

School Of Computing, Informatics, Telkom University, Bandung, Indonesia
Email: [1]fhetetz@student.telkomuniversity.ac.id, [2,*]jondri@telkomuniversity.ac.id, [3]kemasmuslim@telkomuniversity.ac.id
Correspondence Author Email: jondri@telkomuniversity.ac.id
Submitted: **13/08/2023**; Accepted: **25/09/2023**; Published**: 27/09/2023**

Abstract−Online social networking services allow users to post content in the form of text, images or videos. Twitter is a microblogging social networking service that enables its users to send and read text-based messages of up to 140 characters. Retweet is one of the features in Twitter that is important in disseminating information, popular tweets reflect the latest trends on Twitter, the main mechanism that encourages information dissemination is the possibility for users to re-share content posted by their social connections, then it can flow throughout the system. Retweets happen when someone republishes or forwards a post to their homepage and personal profile. Most retweets are credited to the original author of the original post. The retweet prediction system uses an Artificial neural network optimized for Harmony search with tweets about the Jakarta-Bandung Fast Train, which shows the best results when the oversampling method has been carried out with an f1 score of 96.8%.

Keywords: Artificial Neural Network; Harmony Search; Twitter; Retweet; Prediction

1. INTRODUCTION

Media platforms are now relevant and have become part of daily human's life[1], along with technological developments, especially social media. Making a lot of information very fast in the process of spreading and easily accepted by social media users. It's been 7 years to be exact January 21st 2016 when the Jakarta-Bandung High Speed Train was inaugurated by President Joko Widodo, after the inauguration lots of public opinion on various social media platforms, one of which is Twitter.

Information diffusion or dissemination of information is currently greatly increased by the existence of various media platforms social. Twitter is a social media and microblogging service that allows its users to send messages in real time[2]. In Twitter we can provide the widest possible information, users can broadcast their status and rebroadcast someone's status or it is called retweet[3]. Twitter is becoming known for its focus on data analysis, developing marketing and creating a social environment[4], In Twitter social media there is a retweet feature where this feature can speed up the process of spreading information, the retweet process occurs when a user sees post content then there is a desire to spread information such[5].

Artificial neural networks are part of artificial intelligence. Artificial Neural networks are believed to have quite effective approaches such as knowing a pattern, classification, prediction, clustering and forecasting time series with a high accuracy value [6].

Harmony search is an evolutionary machine learning that can be used for optimization problems. Harmony search is an algorithm inspired by musical performances when musicians search for better harmonies. The search for harmony in the process of musical improvisation aims to get the best state based on aesthetic estimates. With this analogy, HS performs an optimization process to obtain the state best[7].

In 2019 Eko Prasetiyo Rohmawan conducted research "On Time Prediction of Student Graduation Using the Decision Tree Method and Artificial Neural Network". In their research, the authors compared the two methods, namely the Decision Tree and Artificial Neural Network, from the results of this study it was proven that the Artificial Neural Network method was better than the Decision Tree method, this was because the data used in the research was in the form of label data with an accuracy value for the Artificial Neural Network of 79.74% [8].

In 2020 Inaaratul Chusna Ichda and her partners conducted research with the title "Optimization of Scheduling Production Using the Harmony Search Algorithm Approach at PT Adi Satria Abadi". In their research, the results obtained using the company method have an average makespan of 0.9 months, whereas if you use the Harmony Search algorithm, you get an average result of 0.8 months. So that the Harmony Search algorithm method can minimum of 0.1 months[9].

Many studies regarding retweet prediction systems, Hamidan Amarullah and his partners have built a system retweet prediction using artificial neural network. In their research, the dataset used is an imbalance class. The authors compare scenarios with two methods, namely oversampling and undersampling. The results show that using the undersampling method is the best result with f1score results 86%[10].

In 2021 Muhammad Syah Zannuar and his partners conducted research with the title "Retweet Predictions Based on User-Based Features Using the Random Forest Classification Method". In this research, Muhammad Syah Zannuar aims to build a retweet prediction system using the user-retweet feature. based and Random Forest in its implementation with an accuracy of 70% [11].

In 2022 Edvan Tanzul and his partners have built a retweet prediction system using the network artificial neural optimized by genetic algorithm. In their research, they made 2 scenarios, there are without using hyperparameter

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 407−413
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4079

tuning and the second using hyperparameter tuning. The results shown by their research using the split data method get quite good performance results, there is content-based accuracy value of 90.4% and user-based 69.11% [12].

In this study, the authors built a retweet prediction model, the data taken to perform This retweet prediction is taken from the Twitter social media platform using the Twitter API, the data obtained is 2204 data. The features used in building this retweet prediction are User-Based and Content-Based features, the method used to build the system is an Artificial Neural Network optimized by Harmony Search algorithm.

2. RESEARCH METHODOLOGY

2.1 Research Stages

This research will build a retweet prediction system using an Artificial Neural Network optimized with Harmony Search as performance. The dataset was obtained through the Twitter API using Netlytic to get 2204 data.

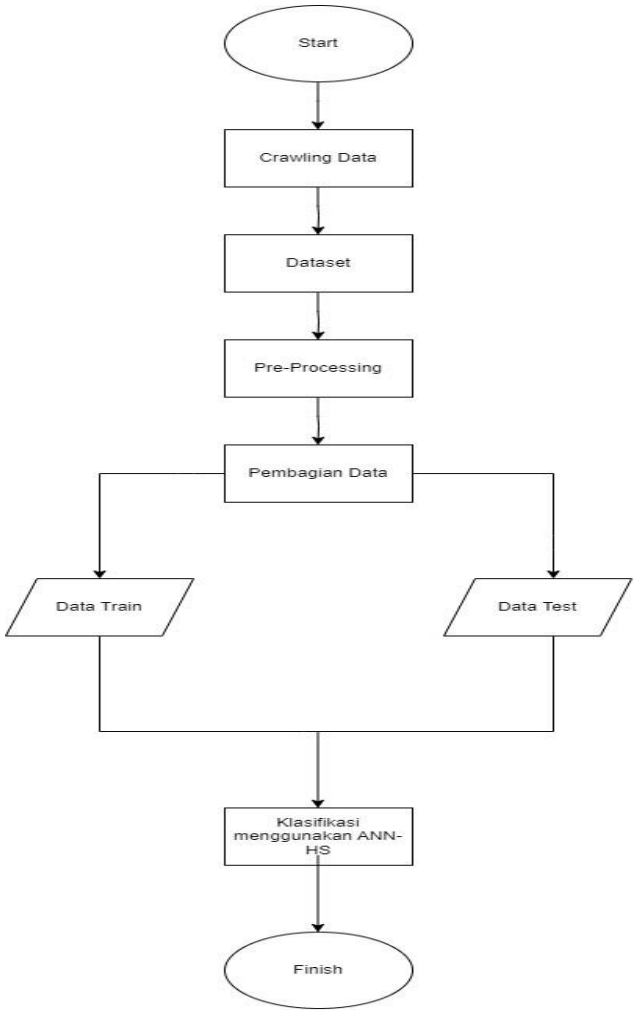

Figure 1. System Design

In the flowchart above in Figure 1, the first stage is crawling data using Netlytic as much as 2204 data. In the second stage, pre-processing is carried out where at this stage data duplication is checked and class disappointment is checked on the data that has been obtained which aims to convert raw data into a form that is easy to understand. In the third stage, data sharing was carried out in this study using ratios of 70/30, 80/20 and 90/10. The final stage is classification using Artificial Neural Network-Harmony Search and conducting evaluations to measure performance.

2.2 Related Studies

Many studies regarding retweet prediction systems, Hamidan Amarullah and his partners have built a system retweet prediction using artificial neural network. In their research, the dataset used is an imbalance class. The authors compare scenarios with two methods, namely oversampling and undersampling. The results show that using the undersampling method is the best result with f1score results 86%[10].

Another study was conducted by Muhammad Rizqi Akbar using the Ensemble Stacking method. In his research Muhammad Rizqi used the Ensemble Stacking method through the K-fold Cross Validation process,

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 407−413
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4079

Ensemble Stacking was formed with 3 base-learners namely Random Forest, Gradient Boosting and Support Vector Machine. The modeling that has been done by the researcher gets the best results when imbalance class handling has been carried out using the smote technique with an F1-score of 86.46%[13].

Another study was conducted by Tasya Maula using the Naïve Bayes classification method. In his research, the aim is to create a retweet prediction system and see how the performance is generated by the Naïve Bayes classification using the User-Based and Content-Based features. The results obtained from his research using k-fold cross validation as a dataset division with a value of k = 10 get an F1-Score of 86.42%[14].

2.3 Dataset

Can be seen in table 1, tweet data is obtained through a crawling process on the Twitter application using the Netlytic platform. The data obtained was 2,204 tweets in Indonesian with the keyword "Jakarta-Bandung Fast Train", using the User-Based and Content-Based features.

Table 1. Retweet features

Feature Type	Feature Name	Data Type	Description
User Based Features	Total_of_tweet	Numerical	The number of previously posted tweets by the user
	No_of_followers	Numerical	The number of people who follow the user
	Age_of_account	Numerical	The number of days since the account was created
	No_groups_user	Numerical	Number of user groups or communities
	Account_verified	Numerical	Confirm account authenticity
	Aver_favou_per_day	Numerical	Average obtained by dividing between Total_of_tweets and Age_of_user
Content Based Features	Contain_location	Boolean	Tweets containing the location name
	Contain_rt_term	Boolean	Tweets containing the word “RT”
	Opt_length	Boolean	Content length between 70 and 100
	Sentimen_Level	(Positive,Negative,Objective)	Tweets classified by sentiment level

2.4 Data Processing

The data has an imbalance class shown in Figure 2, data that has been obtained through Netlytic as many as 2204 data has gone through the preprocessing stage such as data cleaning and checking duplicate data so that the data becomes 1844 data, after going through the cleaning and duplication checking stages then checking for unbalanced classes where class 0 is the class that contains tweets that are not retweeted while for class 1 tweets that are retweeted, where class 0 has 1793 data while for class 1 has 51 data. This problem is called an imbalance class, the handling that is done is to do oversampling so that the data becomes evenly distributed.

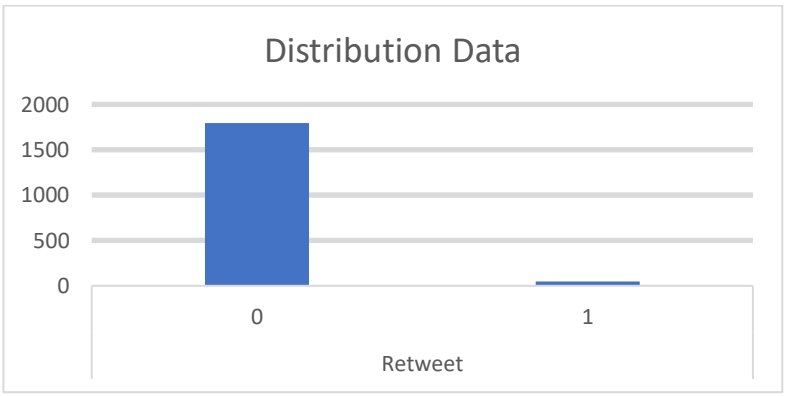

Figure 2. Class retweet distribution

2.4 Artificial Neural Network

Artificial Neural Network (ANN) is inspired by the awareness of the complex learning system in the brain which consists of sets of neurons that are closely interconnected. Neuron networks are capable of performing very complex tasks such as classification and pattern understanding. Artificial Neural Networks show an effective approach for general purposes to find out patterns, classification, clustering and especially forecasting time series with a high degree of accuracy[15]. Classification using ANN requires 3 main stages, there are the input layer as the input data source, the hidden layer as the process and the output layer as the end result of the classification[15]. As seen in Figure 3 below :

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 407−413
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4079

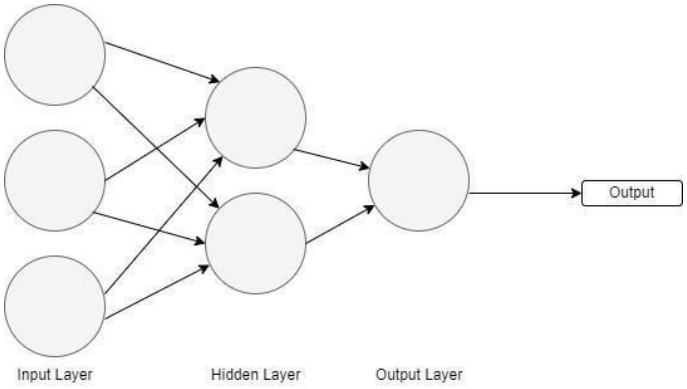

Figure 3. Artificial Neural Network Architecture

The input layer is part of the data input that will be classified to obtain patterns and models, the hidden layer is the hidden layer that receives data from the input, while the size of the input value depends on the multiplication of the weight values [16]. The resulting data weight values will be forwarded to the output layer stage. The working principle of this output layer is almost the same as the hidden layer principle, there is using the sigmoid function before the classification results come out [16].

2.5 Harmony Search

The HS (Harmony Search) algorithm is one of the optimization algorithms that can be used for decision making which is a metaheuristic method inspired by jazz composition[17]. Like a musician who plays certain music, improvises playing notes randomly or based on experience to find beautiful harmony, variables in Harmony Search have random values or values obtained from iteration (memory) in an effort to find optimal solutions. This is then likened to an optimum solution of an optimization problem in terms of retweet prediction [18]. By using the Harmony Search algorithm, it is hoped that optimal retweet predictions can be obtained. In general, the steps to find the optimal optimal solution using the Harmony Search Algorithm are divided into several stages, including :

a. Initialize problem parameters and algorithm.
b. Initialize harmony memories, create harmony memories and assign random sources to each memory.
c. Improvise new harmonies, generate new harmonies based on parameters defined by the user.
d. Update harmony memory, if new harmony is better than worst harmony in harmony memory then replace worst harmony with new harmony.
e. Check the termination criteria, if the termination criteria are met, return the best harmony as the optimal solution for the given problem. If not, return to point 3.

2.6 SMOTE (Synthetic Minority Oversampling Technique)

SMOTE (Synthetic Minority Oversampling Technique) is a commonly used oversampling method when solving problems such as imbalanced data distribution. SMOTE (Synthetic Minority Oversampling Technique) has the goal of balancing class distribution by increasing the number of minority data randomly by creating synthesis data for oversampling purposes[19].

2.7 Confusion Matrix

Confusion matrix is one method that can be used to measure the performance of a classification method. Basically the confusion matrix contains information that compares the classification results performed by the system with the classification results that should be [20].

Table 2. Confusion Matrix

		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP	FP
	Negative	FN	TN

a. Accuracy
Accuracy is the calculation of the actual data from the entire amount of data. Accuracy calculations can be seen in equation 1 below :

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

b. Precision
Precision is the level of accuracy between the information requested by the user and the answers given by the system. The calculation of precision can be seen in the following equation 2 :

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 407−413
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4079

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

c. Recall
Recall is the success rate of the system in retrieving information. The recall calculation can be seen in equation 3:

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

d. F1-score
The F1-Score is a system performance measurement that combines precision scores with recall. Calculation F1-Score can be seen in equation 4 :

$$F1 - Score = \frac{2TP}{2(TP+FP+FN)} \quad (4)$$

3. RESULT AND DISCUSSION

Implementation of Artificial Neural Network with harmony search algorithm can improve the performance of the model by finding the optimal parameter configuration for artificial neural networks. Harmony search is a metaheuristic algorithm inspired by the process of finding harmony in music. It can be applied to optimize various problems, including setting parameters in artificial neural networks. This study uses a dataset consisting of 2204 collected data, the dataset includes two features: User-Based and Content-Based. Evaluation was carried out to determine the best combination results, in this study three different test scenarios were carried out.

a. Testing Scenario 1
In the first test, the dataset used has not been oversampled so it is still in an imbalance class state.

b. Testing Scenario 2
In the second test, the dataset used has been oversampled to avoid class imbalances, the data that has been oversampled can be seen in Figure 4.

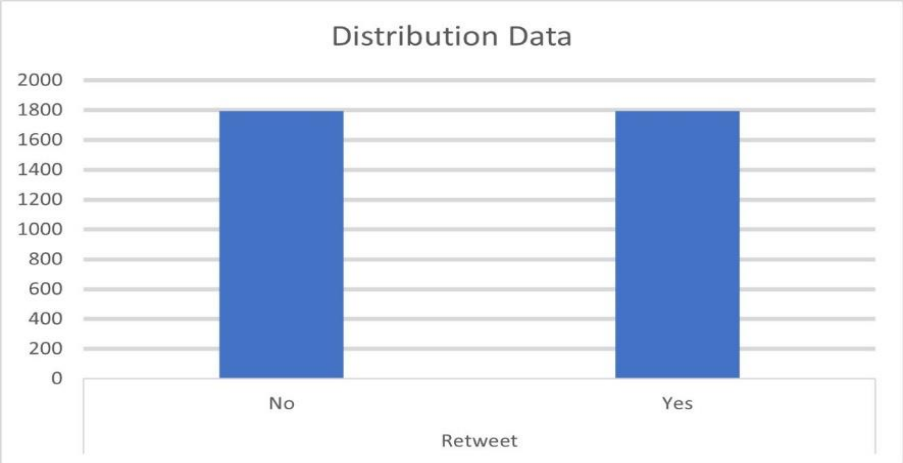

Figure 4. Data that has been oversampled

c. Testing Scenario 3
In the third test, the dataset that has been oversampled is then added using hyperparameter tuning, hyperparameters and values used for tuning can be seen in table 3 :

Table 3. Hyperparameter that has been tunned

Parameter	Value
Learning rate	(0.001, 0.1)
Batch size	(32, 128)
Number of hidden units	(64, 256)
Dropout rate	(0.1, 0.5)

3.1 Results of Scenario 1 Testing

In the first test, the dataset used is an imbalanced class so that oversampling has not been done. The accuracy results obtained using the Artificial Neural Network-Harmony Search method with a ratio of 80/20 for User-Based features is 97.5% and for Content Based 74.8%. This result is considered not good because there are still 0 values for precision, recall and f1-score because the dataset is in the form of an imbalanced class, complete results can be seen in table 4.

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 407−413
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4079

Table 4. Result of Scenario 1 Testing

Features	Acc	Precision	Recall	F1 - Score
User-Based	97.5%	0	0	0
Content-Based	74.8%	0	0	0

3.2 Results of Scenario 2 Testing

In the second test, oversampling has been carried out so that the dataset is not in an imbalanced class state. The results given are better than scenario 1 by using an 80/20 ratio for User-Based to get an f1-score of 98.4% while for Content-Based to get a value of 68.9%, complete results can be seen in table 5.

Table 5. Result of Scenario 2 Testing

Features	Acc	Precision	Recall	F1-Score
User-Based	98.4%	97.3%	99.5%	98.4%
Content-Based	59.5%	54.6%	93.5%	68.9%

3.3 Results of Scenario 3 Testing

In the third test, oversampling has been carried out so that the dataset is not in an imbalance class state and hyperparameter tuning has been added. In testing scenario 3 using 3 ratios, namely 70/30, 80/20 and 60/40, the results given show that the User-Based feature gets the best results at the 80/20 ratio with an f1-score value of 99.8%, while for the Content feature -Based gets the best score when the ratio is 70/30 with an f1-score of 70.1%. complete results can be seen in table 6 and 7

Table 6. Result of Scenario 3 User-Based Features

Method	Acc	Precision	Recall	F1 - Score
70/30	99.7%	99.6%	99.8%	99.7%
80/20	99.8%	99.7%	100%	99.8%
60/40	99.5%	99%	100%	99.4%

Table 7. Result of Scenario 3 Content-Based Features

Method	Acc	Precision	Recall	F1 - Score
70/30	62.2%	57.8%	89%	70.1%
80/20	61.9%	57.3%	88.4%	69.5%
90/10	62.1%	56.8%	88.1%	69%

4. CONCLUSION

In the research that has been done, it can be concluded that the retweet prediction regarding the tweet "Bandung-Jakarta fast train" using the Artificial Neural Network-Harmony Search method will get an increase in performance when using Hyperparameter Tuning. With the use of this Hyperparameter, the Artificial Neural Network method has succeeded in increasing the F1-Score value with a ratio of 80/20 with a result of 99.8% and an accuracy of 99.8%. For further research, the Artificial Neural Network method can be developed with a combination of other optimization algorithms. In addition, the authors also suggest adding other features.

REFERENCES

[1] B. Robert and E. B. Brown, *SOCIAL MEDIA AND SOCIAL ORDER*, no. 1. 2021.
[2] C. Setiawan, “Obesitas , Olahraga , dan Diet : Analisis Sentimen pada Twitter Berbasis Analitik Big Data,” no. March 2022, pp. 71–81, 2023.
[3] F. Zahria Emeraldien, R. Jefri Sunarsono, and R. Alit, “Twitter Sebagai Platform Komunikasi Politik Di Indonesia,” *J. Teknol. dan Inf.*, vol. 14, no. 1, pp. 21–30, 2019, [Online]. Available: www.statisticbrain.com
[4] R. Mchaney and D. D. Sacht, *Web 2.0 and Social Media*. 2016.
[5] R. H. Anggia, Jondri, and K. M. L, “Prediksi Retweet Berbasis Fitur Content Similarity dan Content Based Dengan Menggunakan Metode Support Vector Machine (SVM),” vol. 8, no. 5, pp. 11164–11173, 2021.
[6] B. Y. Pandji, I. Indwiarti, and A. A. Rohmawati, “Perbandingan Prediksi Harga Saham dengan model ARIMA dan Artificial Neural Network,” *Indones. J. Comput.*, vol. 4, no. 2, pp. 189–198, 2019, doi: 10.21108/indojc.2019.4.2.344.
[7] L. Abualigah, A. Diabat, and Z. W. Geem, “A comprehensive survey of the harmony search algorithm in clustering applications,” *Appl. Sci.*, vol. 10, no. 11, pp. 1–26, 2020, doi: 10.3390/app10113827.
[8] E. P. Rohmawan, “Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Decision Tree dan Artificial Neural Network,” *J. Ilm. MATRIK Vol.20 No.1, April 201821-30*, vol. 20, no. 1, pp. 21–30, 2018.
[9] A. Harmony, S. Di, P. T. Adi, and S. Abadi, “OPTIMASI PENJADWALAN PRODUKSI MENGGUNAKAN PENDEKATAN Jurnal DISPROTEK,” vol. 11, pp. 7–12, 2020.
[10] H. Amarullah Purwaatmaja Ash-Shidiq EFSA and K. Muslim Lhaksmana, “Prediksi Retweet Menggunakan Fitur Berbasis

Building of Informatics, Technology and Science (BITS)
Volume 5, No 2, September 2023 Page: 407−413
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i2.4079

Pengguna dan Fitur Berbasis Konten dengan Metode Klasifikasi ANN," vol. 8, no. 5, pp. 11174–11182, 2021.
[11] R. Rakes, J. Jondri, and ..., "Prediksi Retweet Berdasarkan Feature User-based Menggunakan Metode Klasifikasi Support Vector Machine," *eProceedings ...*, vol. 8, no. 5, pp. 11183–11191, 2021, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/15630%0Ahttps://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/15630/15343
[12] T. Akhir, "Prediksi Retweet Menggunakan Fitur User based dan Content Based dengan Metode Klasifikasi ANN-GA Program Studi Sarjana Informatika Fakultas Informatika Universitas Telkom Bandung," 2022.
[13] M. R. Akbar *et al.*, "Prediksi Retweet Berdasarkan User-Based dan Content - Based Menggunakan Metode Ensemble Stacking," vol. 10, no. 2, pp. 1950–1962, 2023.
[14] H. A. P. A.-S. EFSA, Jondri, and K. M. Lhaksmana, "Prediksi Retweet Menggunakan Fitur Berbasis Pengguna dan Fitur Berbasis Konten dengan Metode Klasifikasi ANN," vol. 8, no. 5, pp. 11207–11215, 2021.
[15] M. S. A. Hapsary, S. Subiyanto, and H. S. Firdaus, "Analisis Prediksi Perubahan Penggunaan Lahan Dengan Pendekatan Artificial Neural Network Dan Regresi Logistik Di Kota Balikpapan," *J. Geod. Undip*, vol. 10, no. 2, pp. 1–10, 2021.
[16] R. Zannah, "Analisis Sentimen Pada Media Sosial Twitter Untuk Klasifikasi Opini Islam Radikal Menggunakan Jaringan Saraf Tiruan," no. September 2017, pp. 46–54, 2019, [Online]. Available: http://digilib.uinsby.ac.id/32982/
[17] W. Hartawan, "Otomasi Pid Tuning Untuk Optimasi Kontrol Quadcopter Menggunakan Metode Harmony Search," *J. Inov. Tek. Inform.*, 2021, [Online]. Available: http://journal.universitaspahlawan.ac.id/index.php/jiti/article/view/2012
[18] A. Rahman, E. M. Yuniarno, and I. K. E. Purnama, "Optimasi Penjadwalan Perkuliahan Menggunakan Metode Harmony Search," *Al-Khwarizmi J. Pendidik. Mat. dan Ilmu Pengetah. Alam*, vol. 2, no. 2, pp. 47–58, 2018, doi: 10.24256/jpmipa.v2i2.111.
[19] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, "Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 21, no. 3, pp. 677–690, 2022, doi: 10.30812/matrik.v21i3.1726.
[20] N. Hadianto, H. B. Novitasari, and A. Rahmawati, "Klasifikasi Peminjaman Nasabah Bank Menggunakan Metode Neural Network," *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 163–170, 2019, doi: 10.33480/pilar.v15i2.658.