e-ISSN: 2460-870X | p-ISSN: 2085-1367

Retweet Prediction Using ANN Method and Artificial Bee Colony

Kamaludin Hanif Farisi[a,1], Jondri[b,2], Kemas Muslim L[c,3]
[1,2,3] School of Computing, Telkom University Bandung, Indonesia
e-mail: *kamaludinhaniffarisi@student.telkomuniversity.ac.id[1], [2]jondri@telkomuniversity.ac.id, [3]kemasmulsim@telkomuniversity.ac.id

ABSTRACT

In the ongoing modern era, the rapid dissemination of information takes place, utilizing various channels for data exchange. One such platform is the social media platform Twitter, renowned for its swift and extensive information propagation. A pivotal factor contributing to information distribution on Twitter is the retweet feature, whereby users can redistribute content to their audience. A study has been conducted to forecast this retweet activity by employing the Artificial Neural Network classification method in conjunction with the Artificial Bee Colony optimization approach. The data utilized for this study was collected using data crawling techniques via the Netlytic website, with a total of 2500 pieces of data. Data preparation, model development, and testing in this study are executed using Google Colab. This study leverages diverse features, encompassing content-based feature, user-based feature, and time-based feature. The results from this study reveal that the proposed method achieves an accuracy value of around 83% with the highest accuracy value reaching 84.40%. Based on the obtained results, the Artificial Neural Network (ANN) method optimized by the Artificial Bee Colony algorithm performs better compared to the standalone ANN method. These findings indicate that the fusion of the Artificial Neural Network classification method executed with optimization using the Artificial Bee Colony algorithm yields dependable and consistent performance in predicting retweet activities. For future research, it is expected to be able to develop the ANN method by combining with other optimization algorithms to obtain better results.

Keywords : Artificial Bee Colony, Artificial Neural Network, Prediction, Retweet, Twitter

Article Info :
Submitted:10 Mei *Reviewed: 30 Mei* *Accepted :15 June*

Copyright © 2022 – CSRID Journal. All rights reserved.

1. INTRODUCTION

The dissemination of information in the current modern era is extremely fast and massive, coupled with the rapid development of technology. Nowadays, obtaining and spreading information is very easy due to the abundance of media and platforms available. One of the widely used information-sharing media is social networking platforms, with Twitter being a prominent example.

Twitter is a highly popular microblogging social media platform established in 2006 [1]. It plays a significant role as one of the main channels for creating and disseminating news. On Twitter, users can write tweets containing text, images, and videos. Each tweet can be shared by other users through a feature called 'retweet.' The purpose of the retweet is to enable users to share a tweet from someone else with their followers on Twitter, aiming to help them gain access to the information [2]. Usually, retweets happen when a user likes or finds a tweet interesting. The user then desires to share it with their followers. This retweet process continues and accumulates as more users share the tweet. A tweet becomes very popular and widely discussed if it receives a significant number of retweets, replies, and likes—the more retweets, replies, and likes, the more popular the tweet becomes.

However, with the vast amount of information and tweets circulating on Twitter, not all tweets can become popular, and some may not receive any retweets at all. This raises the question of how a tweet can become popular and how it can influence and captivate an audience. Therefore, this study aims to analyze how a tweet can achieve popularity, whether it will be shared or retweeted by other users, and whether we can predict the dissemination of a tweet.

Several studies on retweet prediction have been conducted using different methods, and the outcomes of each study vary. One such study was carried out in 2017 [3], focusing on retweet

DOI: https://www.doi.org/10.22303/csrid.1.1.2022.01-10

prediction using the random forest method and achieving reasonably good results. This research also suggested that there are influential factors or features in a retweet prediction system. Then, in 2021, Hamidan [4] conducted a study utilizing the factors or features proposed in the previous research [3], but applied the classification method of Artificial Neural Network, obtaining commendable results with an accuracy and F1 score around 86%. The following year, Edvan Tazul [5] undertook research on retweet prediction using the same classification method, Artificial Neural Network, but this time, the study employed the Generic Algorithm as its optimization method. The combination of ANN and Generic Algorithm produced highly promising results with an accuracy of 90.14%.

Therefore, this study combines the classification method of ANN (Artificial Neural Network) with the optimization algorithm of ABC (Artificial Bee Colony). The ABC optimization algorithm's role is to search for and determine the best architectural model for the constructed ANN model.

Based on the description, the topic addressed in this research is the prediction of retweets using the classification method of Artificial Neural Network, optimized through the implementation of the Artificial Bee Colony algorithm. The research scope encompasses tweet data containing the keyword "pemilu" (election) within the timeframe of July 2023.

The objective set for this research is to design a retweet prediction system using the Artificial Neural Network method, optimized through the Artificial Bee Colony algorithm, to obtain commendable performance and accuracy results.

2. RESEARCH METHODOLOGY

In this research, the method used is the classification method of Artificial Neural Network (ANN) with the optimization method of Artificial Bee Colony (ABC). The data utilized in this study consists of a collection of tweets obtained and gathered through Twitter API using data crawling methods via the Netlytic website, with a total of 2500 data. After the classification process, an evaluation is conducted to assess the system's performance and accuracy. All stages of this research, starting from data pre-processing to evaluation, are executed using Google Colab tools. The system's design flow in this research can be observed in Figure 1.

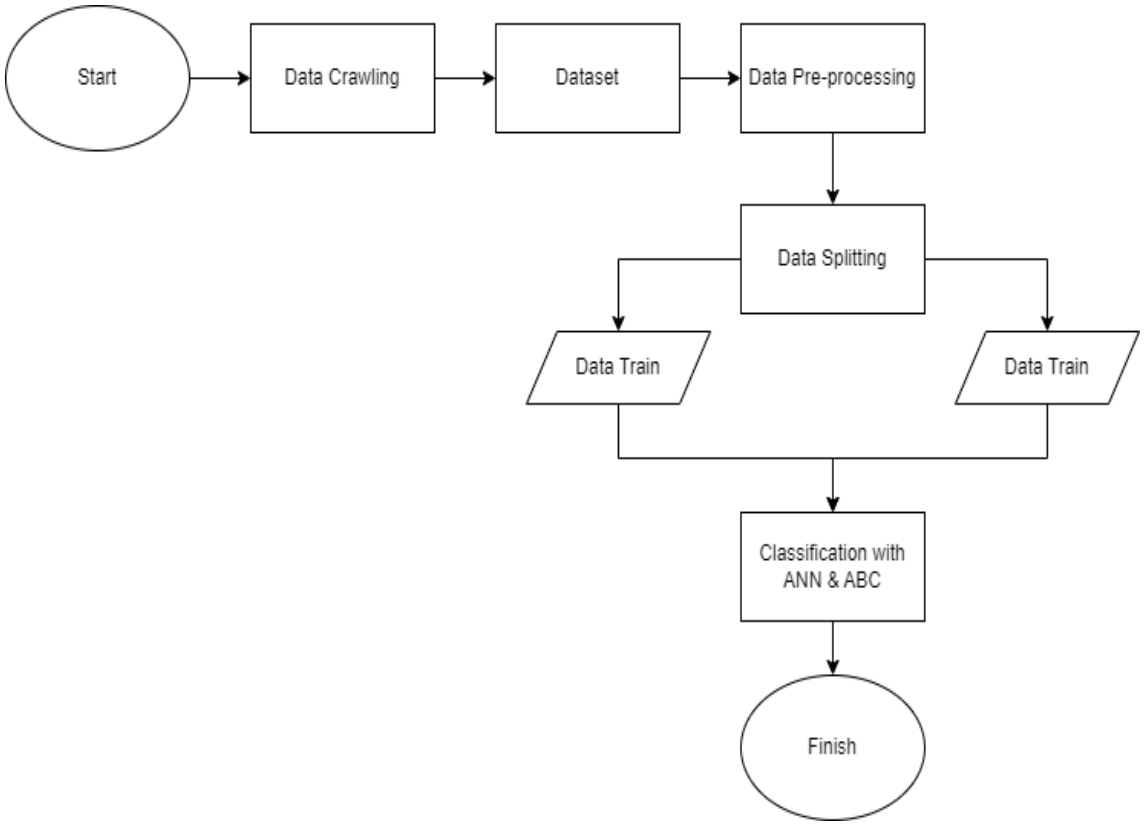


Figure 1. System Design Flowchart

e-ISSN: 2460-870X | p-ISSN: 2085-1367

A. Data Crawling

This research utilizes tweet data retrieved and collected from Twitter API using an authorized token, employing data crawling through the Netlytic website service. From this data collection process, a total of 2500 tweets were gathered, all containing the keyword "pemilu" (election), within the time range of July 13, 2023, to July 14, 2023.

B. Data Pre-processing

The previously collected data, totaling 2500, is in raw form and not ready yet for testing in the model. Hence, this raw data undergoes processing to prepare it for use during testing. Several techniques are employed to process this raw data, namely data cleaning to check for empty or null data and removing duplicate entries. Next, data normalization using MinMaxScaler is applied to each numerical feature to facilitate the classification process. Subsequently, labeling is carried out to introduce features not present in the dataset, such as the "status_retweet" feature containing data represented by 0 for tweets without retweets and 1 for tweets with retweets. The class distribution within "status_retweet" is balanced, with 1267 instances labeled as 0 and 1233 instances labeled as 1. As a result, there is no need to apply the imbalance class technique to the data.

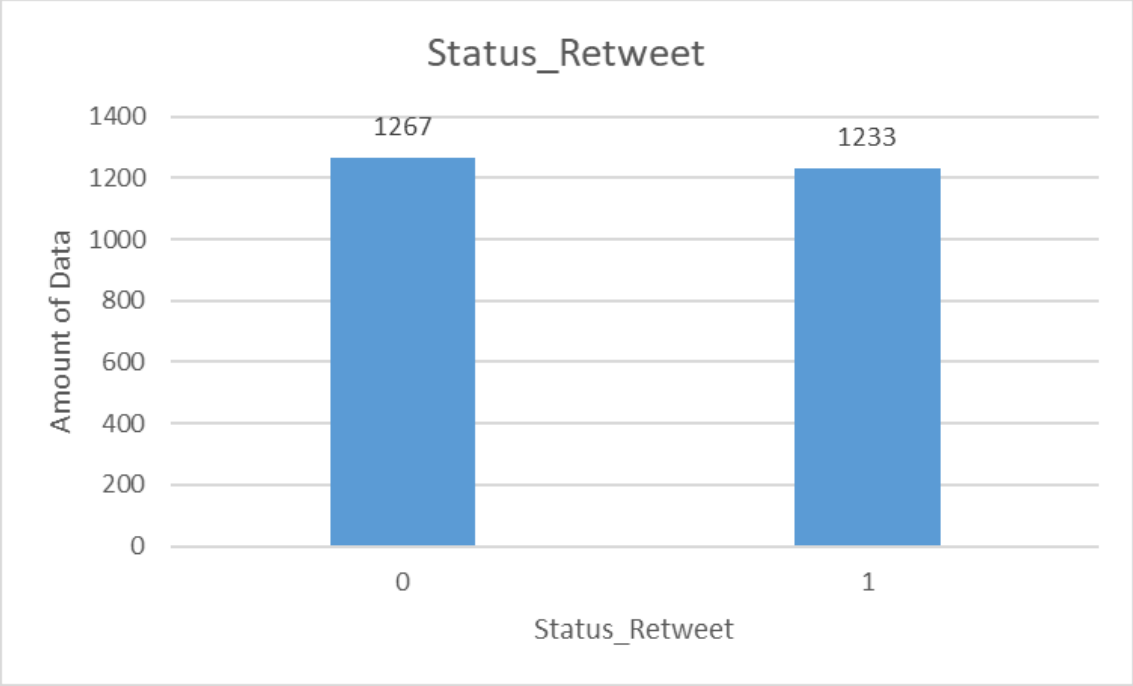


Figure 2. status_retweet Class Distribution

C. Features

The features used in this testing are 26 features with data types of integer and float64. The list of features and their data types is presented in Table 1.

DOI: https://www.doi.org/10.22303/csrid.1.1.2022.01-10

Table 1. List of Features

Nama Fitur	Tipe Data
Total_of_tweets	Int
No_of_followers	Int
No_of_followees	Int
Age_of_account	Int
No_of_favourite	Int
Aver_favou_per_day	Float64
Aver_tweets_per_day	Float64
User_name_len	Int
Is_post_at_hol	Int
Is_post_at_noon	Int
Is_post_at_eve	Int
Is_post_at_weekend	Int
Contain_location	Int
Contain_tvshow	Int
Sentiment_level	Int
Contain_video	Int
Contain_picture	Int
Contain_upper	Int
Contain_number	Int
Contain_excl	Int
Con_user_mentioned	Int
Contain_url	Int
Contain_hashtag	Int
Opt_length	Int
Len_of_text	Int
Status_retweet	Int

D. Artificial Neural Network

The term "ANN," stands for Artificial Neural Network, signifying that ANN is a computational model inspired by the functioning of living organism's neurons [13]. ANN possesses the ability to acquire information and retain it, and it can be defined as a collection of interconnected artificial neuron processing units [13]. Generally, ANN can be divided into three layers: the input layer, the hidden layer, and the output layer [14].

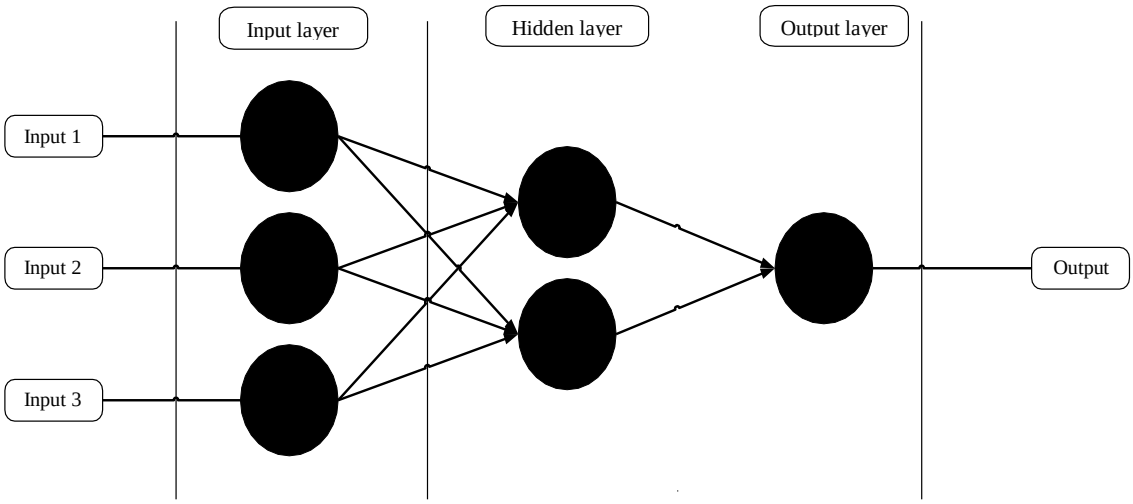


Figure 3 Artificial Neural Network Architecture

E. Articial Bee Colony

Artificial Bee Colony, or ABC, is one of the techniques frequently employed to assist in solving optimization problems due to its relatively simple, rapid, and population-based nature. There are

e-ISSN: 2460-870X | p-ISSN: 2085-1367

three types of bees in the ABC algorithm: worker bees, responsible for finding and obtaining general solutions; observer bees, designed to conduct further exploration to enhance solution accuracy; and scout bees, useful in maintaining the population's search capability [18].

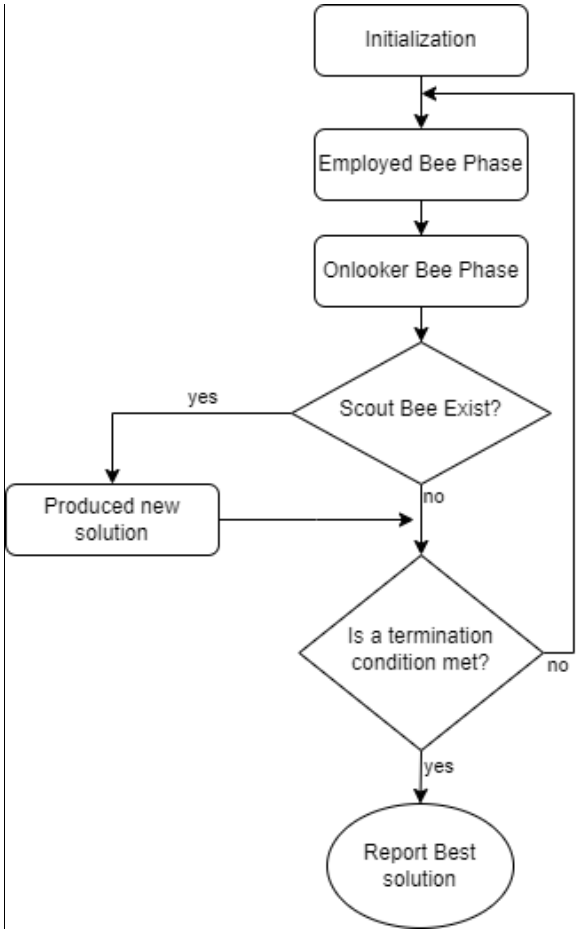


Figure 4 Artificial Bee Colony Flowchart

F. Model and Test Scenario

In this research, testing is conducted using two models for comparison. The first model is the ANN classification model with default input nodes of 26, which is equal to the number of features, a hidden layer size of 30, using the sigmoid activation function, and 100 iterations. While the second model employs the ANN classification model optimized with the Artificial Bee Colony algorithm. ABC functions as an optimization algorithm or hyperparameter tuning technique that works to search for and select the best architecture for the ANN model. ABC algorithm will determine the optimal size of the hidden layer, the best activation function, and the optimal learning rate for the ANN model. The tool used for running the program and conducting testing are Google Colab and Scikit-learn due to its user-friendly nature. The testing scenarios conducted in this study are divided into three categories.

G. Test Scenario 1

Test scenario 1 uses a dataset that has been partitioned into two subsets, namely, the training data and the testing data, with an 80:20 ratio. With a total of 2000 training data, and a total of 500 testing data.

H. Test Scenario 2

Test scenario 2 uses a dataset that has been partitioned into two subsets, namely, the training data and the testing data, with an 70:30 ratio. With a total of 1750 training data, and a total of 750 testing data.

I. Test Scenario 3.

Test scenario 3 uses a dataset that has been partitioned into two subsets, namely, the training data and the testing data, with an 60:40 ratio. With a total of 1500 training data, and a total of 1000 testing data.

J. Confusion Matrix

The confusion matrix is a tabular method used to visualize the performance of a model's predictions. Each entry in the confusion matrix indicates the number of predictions made by the model, classified as either correct or incorrect, based on the respective classes.

Table 2. Confusion Matrix

		ACTUAL	
		POSITIVE	NEGATIVE
PREDICTED	POSITIVE	TRUE POSITIVE	FALSE NEGATIVE
	NEGATIVE	FALSE POSITIVE	TRUE NEGATIVE

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \times 100\% \quad (1)$$

$$Missclassification\ rate = \frac{\text{FP+FN}}{\text{TP+TN+FP+FN}} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (4)$$

$$F1\ Score = 2\ \frac{Recall\ x\ Precision}{Recall+Precision} \quad (5)$$

3. RESULTS AND DISCUSSION

Each test uses a dataset that has been divided into training data and testing data and then tested using the ANN classification model and the ANN classification model optimized with the ABC algorithm, with each model being tested for 100 iterations to find the optimal results. The tool used for performing the testing are Google Colab and Scikit-learn due to its easy to understand.

A. Scenario 1

In the first test, with a data split ratio of 80:20, namely 80% of the data train with a total of 2000 data and 20% of the data test with a total of 500 data. ANN model obtained a good result, where the model achieved a recall value of 81.78%, a precision value of 81.89%, an accuracy value of 83%, and an F1 score of 82.15. %. Then in the second model after optimization, the ABC algorithm finds the best architectural form of the ANN model, namely with hidden layer sizes (39, 27, 34, 44), with a sigmoid activation function, and a learning rate of 0.0074. So that the second model is proven to be able to increase the results of each evaluation score, with the accuracy value increasing to 84.40% and the highest achieved score being precision with a value of 85.96%.

e-ISSN: 2460-870X | p-ISSN: 2085-1367

Table 3. Results of Scenario 1

Model	Recall	Precision	Accuracy	F1 Score
ANN	81.78%	81.89%	83.00%	82.15%
ANN & ABC	84.21%	85.96%	84.40%	83.8%

B. Scenario 2

In the second test using a dataset that has been divided into 70:30, with 70% train data which is 1750 amount of data and 30% test data which is 750 amount of data, the ANN model also gets good results, with the highest score obtained being a recall value of 83.51% and being able to get an accuracy value of 82.53%. In the second test scenario the ABC optimization algorithm finds the best ANN architecture which is different from the first test scenario, where the size of the hidden layer is (17, 16, 44, 21), with a sigmoid activation function, and a learning rate of 0.0063. The second model gets better performance with the greatest increase in the precision value of 83.93% and obtains an accuracy value of 83.60%.

Table 4 Results of Scenario 2

Model	Recall	Precision	Accuracy	F1 Score
ANN	83.51%	81.53%	82.53%	82.5%
ANN & ABC	84.76%	83.93%	83.60%	84.25%

C. Scenario 3

In the third test the dataset used is a dataset divided into 60:40 with 60% train data which is 1500 amount of data and 40% test data which is 1000 amount of data. The first model gets a good results with an accuracy value of 81.51%. Whereas in the third scenario the ABC optimization algorithm finds the best ANN architecture where the size of the hidden layer is (46, 14, 49, 22), with the rectified linear unit (relu) activation function, and the learning rate is 0.0074. The ABC-optimized model exhibited enhanced performance, achieving the highest accuracy value of 83.79%.

Table 5 Results of Scenario 3

Model	Recall	Precision	Accuracy	F1 Score
ANN	81.14%	80.65%	81.51%	80.91%
ANN & ABC	81.95%	81.29%	83.79%	82.63%

D. Discussion

Based on the outcomes of the testing conducted in each scenario, the employed model exhibited commendable and consistent results. The ANN model gets the best results in the second test scenario with a recall value of 83.51%, and gets the highest accuracy value in the first scenario with a value of 83.00%.

The results of this study prove that the Artificial Bee Colony or ABC optimization algorithm can improve the performance of the ANN model, with an increase value of around 1-4%. The ANN model that has been optimized with Artificial Bee Colony consistently improves the performance results of each evaluation value, where the highest value is obtained in the first test scenario with a precision value of 85.96%, and the highest accuracy value obtained is 84.40% in the first test scenario.

The best test scenario based on the obtained results is Scenario 1. When compared to the other test scenarios, Scenario 1 consistently achieves a better average score. Additionally, the first test scenario also obtains the highest accuracy value among the two test scenarios. Therefore, the test scenario that yields the best results is Scenario 1.

Furthermore, it is also important to include information about which features are significant in the classification process by employing techniques such as information gain and mutual information. This can be utilized to determine which features have an impact on whether a tweet will receive retweets or not.

DOI: https://www.doi.org/10.22303/csrid.1.1.2022.01-10

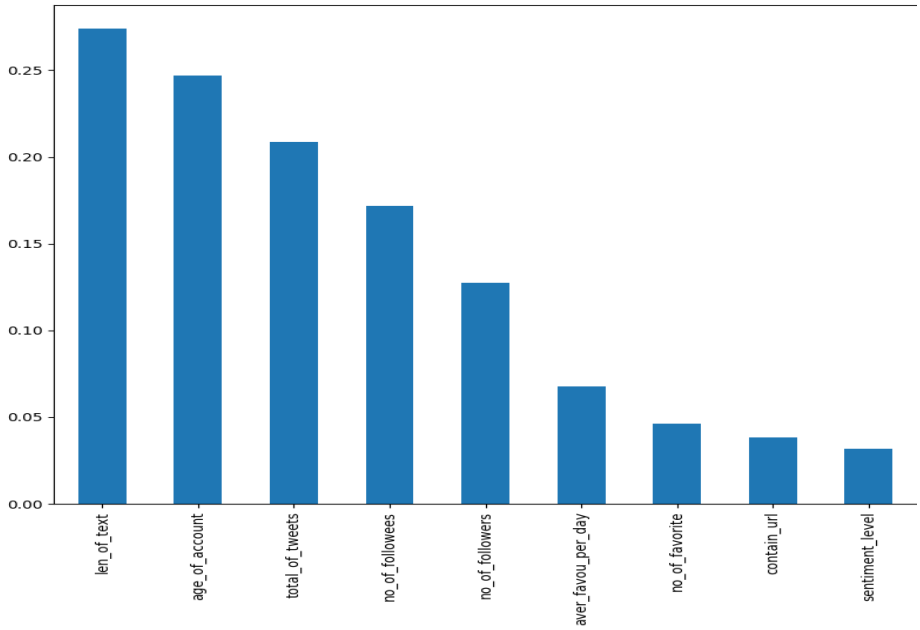


Figure 5 Important Features

The obtained results indicate that each feature has variation information gain scores. As illustrated in Figure 5, the information gain scores have been arranged from the highest to the lowest. Based on the information gain scores, it can be interpreted that "len_of_text" holds the most significant influence in determining the classification. Following this, in order, "age_of_account," "total_of_tweets," "no_of_followers," and "no_of_followees" are also features that significantly impact the classification process.

4. CONCLUSION

In this retweet prediction research using the ANN and Artificial Bee Colony methods, it can be concluded that the constructed model effectively and consistently predicts retweet outcomes. We can see this from the results of each testing phase, where the model consistently attained high and stable values. The application of the Artificial Bee Colony algorithm has demonstrated its capacity to enhance the performance of the ANN classification model. The ANN model that has been optimized with Artificial Bee Colony gets the best accuracy value at 84.40%.

For future studies, it is recommended to explore the potential of enhancing the ANN method through combinations with other methods or different optimization algorithms. Additionally, researchers are advised to consider employing larger datasets, as this would introduce greater data variety to the exploratory process, potentially leading to even more improved outcomes.

e-ISSN: 2460-870X | p-ISSN: 2085-1367

REFERENCES

[1] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a Social Network or a News Media?" in *WWW 2010*, 2010, pp. 591–600.

[2] N. Oliveira, J. Costa, C. Silva, and B. Ribeiro, "Retweet Predictive Model for Predicting the Popularity of Tweets," *Springer Nature Switzerland*, pp. 185–193, 2020, doi: 10.1007/978-3-030-17065-3_19.

[3] J. Mothe, T. B. Ngoc, and Hoang, "Predicting information diffusion on Twitter–Analysis of predictive features," *J Comput Sci*, vol. 28, pp. 257–264, 2018.

[4] 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," in e-Proceeding of Engineering, 2021, pp. 11207–11215.

[5] E. T. Arifin, "Prediction Retweet Using User-Based and Content-Based with ANN-GA Classification Method," *Technology and Science (BITS)*, vol. 4, no. 2, 2022, doi: https://doi.org/10.47065/bits.v4i2.1931.

[6] Wibowo and E. Winarko, "Paper Review: Data Mining Twitter," Maint. Cult. Herit. Through Inf.Technol. a Smart Futur," 2014, pp. 1–10.

[7] T. S. Maulasirri, Jondri, and K. M. Lhaksamana, "Prediksi Retweet Berbasis Konten dan Pengguna dengan Metode Klasifikasi Naïve Bayes," *e-Proceeding Eng*, vol. 8, no. 5, pp. 11175–11181, 2021.

[8] A. Black, C. Mascaro, M. Gallagher, and S. P. Goggins, "Twitter Zombie: Architecture for Capturing, Socially Transforming and Analyzing the Twittersphere," *Proceedings of the 2012 ACM Group Conference*, pp. 229–238, 2012, doi: 10.1145/2389176.2389211.

[9] H. Becker, M. Naaman, and L. Gravano, "Beyond Trending Topics: Real-World Event Identification on Twitter," *Fifth International AAAI Conference*, vol. 5, no. 1, pp. 438–441, 2011, doi: https://doi.org/10.1609/icwsm.v5i1.14146.

[10] G. Pant, P. Srinivasan, and F. Menczer, "Crawling the Web," in *Web Dynamic*, 2003, pp. 153–177.

[11] E. R. Mamuri and B. Setiawan, "Mendeteksi Pesan Berita Palsu (Hoax) pada Twitter dengan Algoritma Adaboost dan ANP," 2019.

[12] J. E. Sembodo, E. B. Setiawan, and A. Baizal, "Data Crawling Otomatis pada Twitter," *Indonesia Symposium on Computing*, pp. 11–16, 2016, doi: 10.21108/INDOSC.2016.111.

[13] I. N. da Silva, D. H. Spatti, R. A. Flauzino, and L. H. B. Liboni, "Chapter 1 Introduction," in *Artificial Neural Networks*, 2017, pp. 3–19. doi: http://dx.doi.org/10.1007/978-3-319-43162-8_2.

[14] I. N. da Silva, D. H. Spatti, R. A. Flauzino, and L. H. B. Liboni, "Chapter 2: Artificial Neural Network Architectures and Training Processes," in *Artificial Neural Netwoks*, 2017, pp. 21–28. doi: http://dx.doi.org/10.1007/978-3-319-43162-8_2.

[15] S. Hassanipour *et al.*, "Comparison of artificial neural network and logistic regression models for prediction of outcomes in trauma patients: A systematic review and meta-analysis," *Elsevier: Injuri*, vol. 50, no. 2, pp. 244–250, 2019, doi: https://doi.org/10.1016/j.injury.2019.01.007.

[16] D. Karaboga, "An Idea based on honey bee swarm for numerical optimazation," *Technical report -TR06*, 2005.

[17] J. C. Bansal, H. Sharma, and S. S. Jadon, "Artificial bee colony algorithm: a survey," *International Journal of Advanced Intelligence Paradigsm*, vol. 5, no. 1, pp. 123–159, 2013, doi: http://dx.doi.org/10.1504/IJAIP.2013.054681.

[18] H. Gao, Y. Shi, and C.-M. Pun, "An Improved Artificial Bee Colony Algorithm with its Application," *IEEE Trans Industr Inform*, vol. 15, no. 4, pp. 1853–1865, 2018, doi: https://doi.org/10.1109/TII.2018.2857198.

[19] E. Rahm and H. H. Do, "Data Cleaning: Problems and Current Approaches," *IEEE Data Eng. Bull*, vol. 23, no. 4, pp. 3–13, 2000.