Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 522−528
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1931

Prediction Retweet Using User-Based and Content-Based with ANN-GA Classification Method

Edvan Tazul Arifin[1,*], Jondri[2], Indwiarti[3]

School of Computing, Informatics, Telkom University, Bandung, Indonesia
Email: [1,*]edvanarifin@student.telkomuniversity.ac.id, [2]jondri@telkomuniversity.ac.id, [3]indwiarti@telkomuniversity.ac.id
Email Author Correspondence: edvanarifin@student.telkomuniversity.ac.id
Submitted:**22/07/2022**; Accepted:**18/08/2022**; Published**: 30/09/2022**

Abstract− Current technological advances have caused rapid dissemination of information, especially on social media, one of which is Twitter. Retweeting or reposting messages is considered an easily available information diffusion mechanism provided by Twitter. By finding out why a user retweets a tweet from another person and by making this prediction we can understand how information diffuses on Twitter. In this study, Artificial Neural Network – Genetic Algorithm is used in the classification process and uses user-based and Content-Based features. Evaluation result obtained in this study are 90% accuracy, 72% precision, 83% recall, and 65% F1-Score value on the model by Oversampling.

Keywords: Twitter; Retweets; ANN-GA; Classification,Oversampling

1. INTRODUCTION

Social networks are increasingly popular as a medium for disseminating information. Based on data from datareportal.com in January 2021 Youtube became the most widely used social media in Indonesia followed by Whatsapp, Instagram, Facebook and Twitter in fifth place.By using these sites, people can share information on various topics according to their likes and interests. Social networking sites such as Facebook and Twitter show tremendous potential to make content instantly popular[8]. Social media is one of the means to express something, channel talent interests, and disseminate information. Because of its convenience in sharing time in real-time [11]. Information dissemination is an advantage that can be utilized for the benefit of a party so that opinions are made based on wishes [15].

Twitter is one of the most popular social media at this time, Twitter users can perform activities such as sharing information in the form of text, video, or images[8]. And there is a retweet feature, where users can retweet or share other users' tweets that they like to share with their followers so that their followers can find out the information they get[9]. Retweet behavior occurs because a user observes the content of the post and then is interested in the content so there is a desire to post again. This retweet process will continue until other users do not spread it again. This behavior is an interesting review material to observe [15]. In November 2021, there were 63 million active users in Indonesia. Twitter is also very popular among developers because it makes it easy to get the data they need. The process of disseminating information on Twitter using the retweet and like features, with the addition of the two numbers, the more information will be spread [2].

Previous research discusses retweet prediction using the Random forest model in its implementation based on user-based features [2]. And in other studies that discuss the prediction of information diffusion on Twitter using several features, namely user-based, content-based, and time-based, and applied to several classification models including Naive Bayes and Artificial Neural Network [1].

In 2017, Thi Bich Ngoc Hoang and Josiane Mothe conducted a study entitled "Predicting Information Diffusion on Twitter - Analysis of predictive features". In the study, Thi Bich Ngoc Hoang and Josiane Mothe conducted research with the aim of predicting whether a post would be forwarded or. In addition, in the study they also aimed to predict how much the post would spread.In their model experiments based on three types of features, namely user-based, time-based and content-based using Random forest algortihm in the model. Where from the experiments run get satisfactory results and have high accuracy [5].

In 2021, Hamidan Amarullah Purwaatmaja Ash-Shidiq EFSA conducted research with the title "Retweet Prediction Using User-Based Features and Content-Based Features with ANN Classification Motode". In the study, Hamidan aims to predict retweets based on user-based and content-based by using ANN classification method in its implementation. In the study, it can be concluded that the ANN classification method has an accuracy rate of 86% [1].

In 2021, Muhammad Syah Zannuar S conducted research with the title "Retweet Prediction Based on User-Based Features Using the Random Forest Classification Method". In this study, Zannuar aims to build a retweet prediction system using user-based features and the Random Forest method in its implementation. In this study, it can be concluded that the Random Forest method has an accuracy rate of 70% [2].

In this study, the authors built a retweet prediction model using 5874 data. the data taken is Twitter data using the Twitter API. The features used are user-based features and cotent-based features. The tweets used are Indonesian tweets. The purpose of this final project is to build a model that can predict the occurrence of retweets from tweets.

BITS is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 522−528
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1931

The method used in this retweet prediction is the Artificial Neural Network - Genetic Algorithm classification method where ANN has good prediction performance, and can handle complex relationships well and high tolerance for noisy data [2]. Genetic Algorithm is used as the weight of ANN because GA has the ability to get the optimal weight.

Purpose of this research is to predict retweets on a tweet with user-based and content-based features. The method used in this research is ANN-GA, and the data used amounted to 5874 data obtained through the Twitter API. The searched dataset uses the keyword "Kereta Api Cepat Jakarta-Bandung". implementation The author's implementation of the topic modeling program uses ANN and GA modeling as weights in ANN with the python programming language.

2. RESEARCH METHODOLOGY

2.1 Research Stages

In this research, the method used is Artificial Neural Network - Genetic Algorithm and evaluates it with accuracy values. The data used is obtained from twitter by crawling data using the tweepy library and the data obtained is 7302 data.

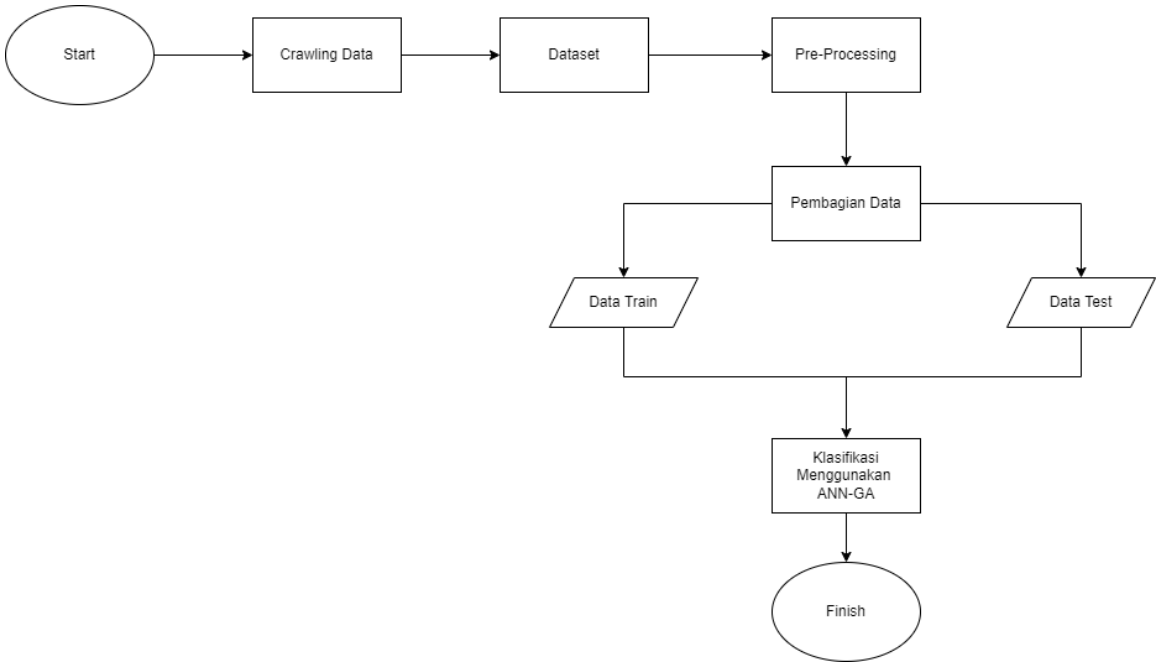

Figure 1. System Design

Judging from the flowchart above, the first stage carried out is crawling data, and retrieving data manually using tweepy and turning it into a dataset that is ready to be processed. The second stage is preprocessing. At this stage, case folding, data cleaning (URL, mention, emojis, digits, punctuation), tokenizing, normalization, stemming / lemmatization. The third stage performs data division. The last stage performs classification and the use of the Artificial Neural Network - Genetic Algorithm method and evaluates to measure the performance of the prediction data trained by the classification model.

2.2 Dataset

In this research, the features used are user-based and content-based features and contain 5847 rows of data.

a. Features *User-Based*
   1. *Total_of_tweet*, is the total of *tweets* previously posted user *the timeline* that has a numeric data type.
   2. *No_of_followers*, is the number of people who follow users who have numeric data type.
   3. *No_of_followees*, is the number of people the user has followed that has a numeric data type.
   4. *Age_of_account*, is the number of days since the account was created that has a numeric data type.
   5. *No_of_favourite*, is the number of tweets that the user likes on the timeline which has a numeric data type.
   6. *No_groups_user*, is the number of groups or communities of user accounts that have a numeric data type.
   7. *Aver_favou_per_day*, is the average obtained from the division between *No_of_followers* and *Age_of_user* which has a numeric data type.
   8. *Aver_tweets_per_day*, is the average obtained from the division between *Total_of_tweets* and *Age_of_user* which has a numeric data type.
   9. *User_name_len*, The length of the user account name that has a numeric data type.

Copyright © 2022 **Edvan Tazul Arifin**, Page 523
BITS is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 522−528
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1931

b. Features *Content-Based*
   1. *Contain_location*, is a *tweet* containing the name of a location that has a Boolean data type.
   2. *Opt_length*, is the content length between 70 to 100 characters which has a Boolean data type.
   3. *Contain_rt_term*, is a *tweet* containing or containing the word "RT" which has a Boolean data type.
   4. *Sentiment_level*, is a *tweet* classified into a sentiment level that has a data type (Positive, Negative, Objective).

2.3 Data Processing

The data obtained is processed first so that it is ready to be used in the model. By cleaning and checking data duplication. Data that was previously 7302 data becomes 5874 data. Then checking the imbalance class on the data obtained, where class 0 is a class that contains tweets that are not retweeted, while class 1 is a class that contains tweets and is retweeted. The data obtained has a class imbalance. The process of handling imbalance classes contained in test scenario 1 is carried out by oversampling using the SMOTE method.

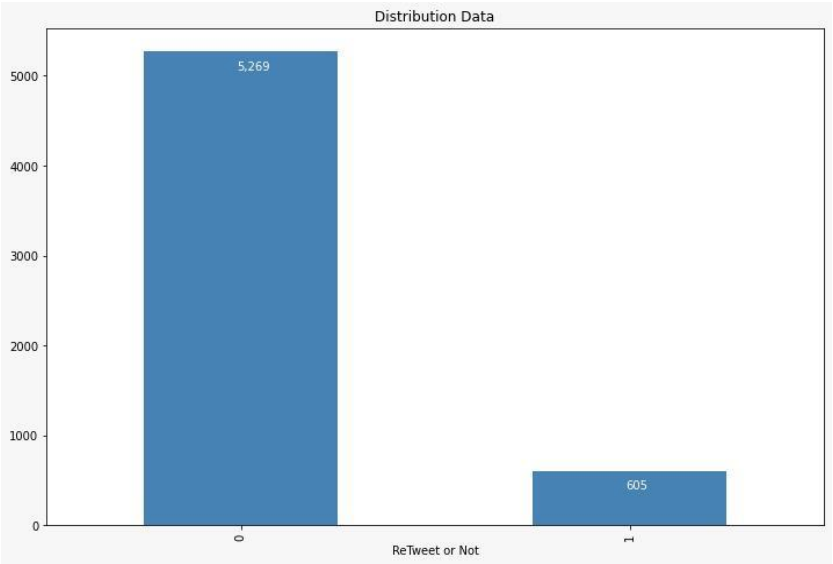

Figure 2. Distribution Class Retweet

2.3 Artificial Neural Network

Artificial Neural Network or ANN is a classification technique for modeling systems that have non-linear and complex properties. The characteristics of this model are determined by the network architecture, activation function, and training process[17]. The network architecture places neurons in a layer consisting of an input layer, output layer, and a hidden layer if any, of which there are one or more [1,12].

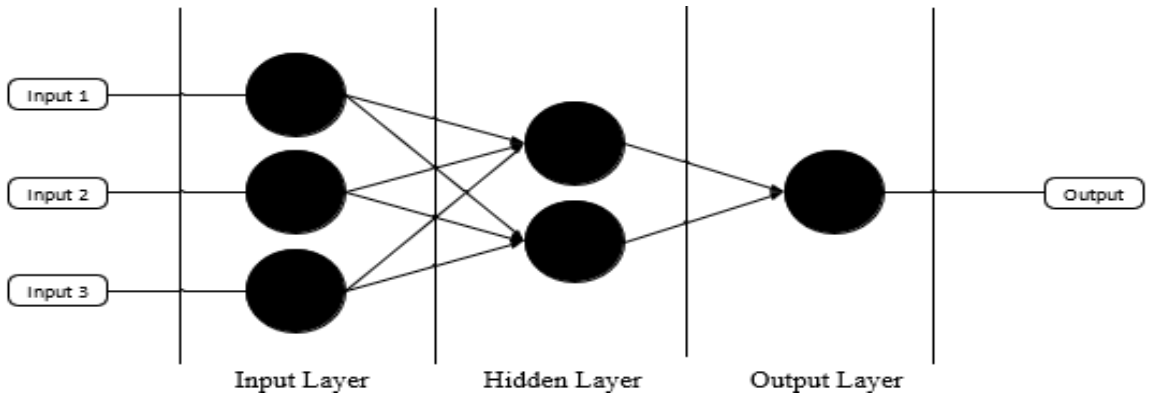

Figure 3. Artificial Neural Network

2.4 *Genetic Algorithm*

Genetic Algorithm is an algorithm inspired by the theory of evolution in an environment where the best individual has a greater chance of survival and a greater chance of spreading the reproductive genome that can produce better offspring than the previous offspring[7] . Genetic Algorithm is also an algorithm that applies the principle of natural selection and has been used successfully in machine learning and optimization problems [3].

Genetic Algorithm has several phases in solving the problems experienced, namely initiation which is the first stage to get new individuals who will later become solutions to problems or programs that are being built[10]. Next

BITS is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 522−528
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1931

grouping, offspring will be grouped into one The next stage is the evaluation stage to calculate individual value in offspring or parent, and finally, selection is a process to select individuals from offspring and parent, the best value will be passed to the next generation or can be a solution [14]. Mutation is a random change that an individuals genes undergo and mutation is the most important factor in evolution which allows you to find new solution[16].

David Goldberg was the first to introduce the genetic algorithm cycle which is described as Figure 2. The cycle starts from creating an initial population randomly, then each individual is calculated for its fitness value. The next process is to select the best individual, then crossover is performed and continued by the mutation process so that a new population is formed. Furthermore, this new population undergoes the same cycle as the previous population. This process continues until the nth generation [4].

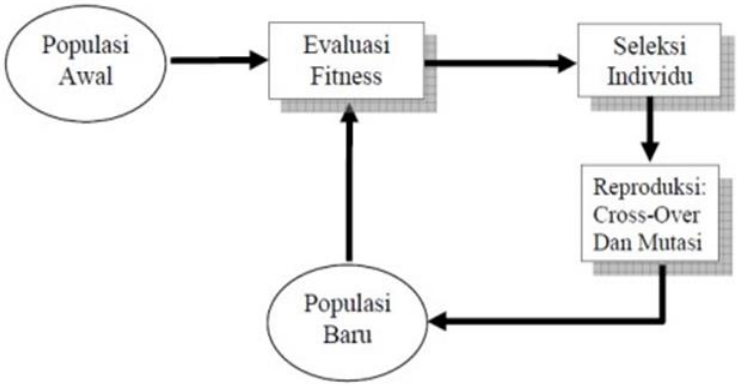

Figure 4. *Genetic Algorithm*

2.5 Confusion Matrix

To determine the performance of the prediction model, measurements were made in several scenarios. Several metrics are used to measure retweets. Evaluation is done using a confusion matrix. The confusion matrix contains classification information in the form of True Positive, True Negative, False Positive, and False Negative[12]. In general, the confusion matrix represented in table 1.

Table 1. *Confusion Matrix*

	Prediction Class	
	True	False
True	TP	TN
False	FP	FN

Based on these values, accuracy, *recall*, precision, and *f-measure* with the following formula:

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

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (2)$$

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

$$F = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

3. RESULTS AND DISCUSSION

In this research using Artificial Neural Network (ANN) and Genetic Algorithm (GA). which is used as the weight of ANN and GA is used in each iteration of ANN. Genetic Algorithm is used as the weight of ANN because GA is able to get the optimal weight, the weight will be entered into the existing layer in ANN. Content-based features used are Contain_location, Opt_length, Contain_rt_term, and Sentiment_level. content-based dataset because it experiences imbalance data, Oversampling is done using SMOTE to balance imbalance data on Content-based. the test scenarios in this study are split datasets and hyperparameters tuning. the two test scenarios will be compared to get the highest value.

3.1 Test Results and Analysis of Scenario 1

In the first test, three splitting experiments were carried out on the dataset, the first division of 70% training data and 30% testing data, the second test of 80% training data and 20% testing data, and the third test of 90% training data and 10% testing data. This test is to determine the results of Accuracy, Precision, recall and F1-Score values on User-

BITS is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 522−528
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1931

based and Content-based datasets. Testing is carried out 10 times for each of the best combination experiments so as to produce the maximum coherence value. Genetic algorithm is used as a weight to produce the best generation obtained.

Table 2. Test Result 1 User-Based

Split	Accuracy	*Precision*	*Recall*	*F1-Score*
70:30	64.95%	62.54%	56.13%	59.16%
80:20	64.09%	67.86%	33.64%	44.97%
90:10	69.11%	72.31%	43.00%	54.02%

It can be seen from the result of the first test user-based datasets in table 2. split datasets do three times split data with 70% data train 30% data testing, 80 data train 20% data testing, 90% data train and 10% data testing get pretty good results, the greatest results are obtained by doing split data 70:30 with F1-Score results 59.16%. while the best accuracy results are obtained with split data 90:10 with 69.11% accuracy results.

Table 3. Test Result 1 Content-Based

Split	Accuracy	*Precision*	*Recall*	*F1-Score*
70:30	90.07%	0.00%	0.00%	0.00%
80:20	89.87%	0.00%	0.00%	0.00%
90:10	90.14%	0.00%	0.00%	0.00%

It can be seen from the results of the first content-based test in table 3, the results obtained by splitting the data 70% data training 30% data testing, 80% data training 20% data testing, and 90% data training 10% data testing. get the best accuracy results of 90.14% by splitting the data 90:10. precision, recall, and F1-Score values from each test get a value of 0.00% because the content-based dataset is imbanaced data, the next content-based test will do Oversampling using SMOTE.

Table 4. Test Result 1 Content-Based using Oversampling

Split	Accuracy	*Precision*	*Recall*	*F1-Score*
70:30	90.07%	83.87%	83.87%	65.44%
80:20	89.87%	80.64%	80.64%	64.82%
90:10	90.14%	47.25%	100%	64.18%

It can be seen from the results of testing content-based datasets using Oversampling in table 4, the results obtained are different from the results in Table 3 datasets before using Oversampling get the results of Precision, Recall, and F1-Score 0 due to dataset imbalance. datasets in table 4 have been balanced with SMOTE. split data is done with 70% data train 30% data testing, 80% data training 20% data testing, 90% data training and 10% data testing get good results with slightly different results. split data model 70% data training and 30% data testing get F1-Score 65.44% results with 90.07% accuracy.

3.2 Results of Testing and Analysis of Scenario 2

In the second test, testing was carried out using Hyperparameters tuning. Hyperparameters tuning is used to influence the learning process and the weight used in this test is Genetic Algorithm because GA is able to get the optimal weight. epochs are used to determine the number of times the algorithm will work, in this study the value of epochs is filled with 5, 10 and 15. batch size is used to determine the number of samples worked on before updating internal model parameters, in this study the value of batch size is filled with 32, and 64. Activation Function is used to introduce nonlinear models and learn the limits of nonlinear predictions, the value of the activation function is filled with sigmoid, relu and softmax. can be seen in table 5 for the hyperparameters tuning scenario that will be used.

Table 5. Skenario Test 2 Hyperparameters Tuning

Hyperparameters	Value
epochs	5,10,15
Batch Size	32,64
units	3,5, 7
Activation Function	Sigmoid, relu, softmax

Table 6. Test Result 2 Content-Based

Activation_Function	Batch_Size	Epochs	Units	Akurasi	Precision	Recall	F1-Score
Sigmoid	32	10	3	56.70%	53.53%	80.21%	64.21%

It can be seen from the results of testing the second content-based scenario in table 6, the results obtained on content-based datasets that have been Oversampling using SMOTE. in this test using Hyperparameters tuning and getting the best results with the Activation Function value getting a sigmoid value, Batch_size with the best value of

BITS is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 522−528
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1931

32, epochs 10, units 3, Accuracy 56.70%, Precision 53.53%, recall 80.21% and F1-Score value 64.21% for content-based datasets using Hyperparameters.

Table 7. Test Result 2 User-Based

Activation_Function	Batch_Size	Epochs	Akurasi	Precision	Recall	F1-Score
Sigmoid	32	15	68.73	71.21%	43.12%	53.71%

It can be seen from the results of testing the second user-based scenario in Table 7, the results obtained on user-based datasets using Hyperparameters tuning get the best results with Activation Function values getting sigmoid values, Batch_size with the best value of 32, epochs 15, Accuracy 68.73%, Precision 71.21%, recall 42.12% and F1-Score value 53.71% for user-based datasets using Hyperparameters.

3.3 Test Results Analysis

Based on the results obtained from the first and second tests, it can be seen in Figure 5 that the influence caused by the features in the data from the model that has been built. Verified has the highest influence value on a tweet to be retweeted, followed by No_of_followers which is the number of followers of the user account in second place after verified, then followed by Aver_favou_per_day which is the average obtained between Total_of_tweets and Age_of_user in third place. Aver_tweetz_per_day which is the average between Total_of_tweets and Age_of_user has no influence with the smallest ratio of the model that has been built.

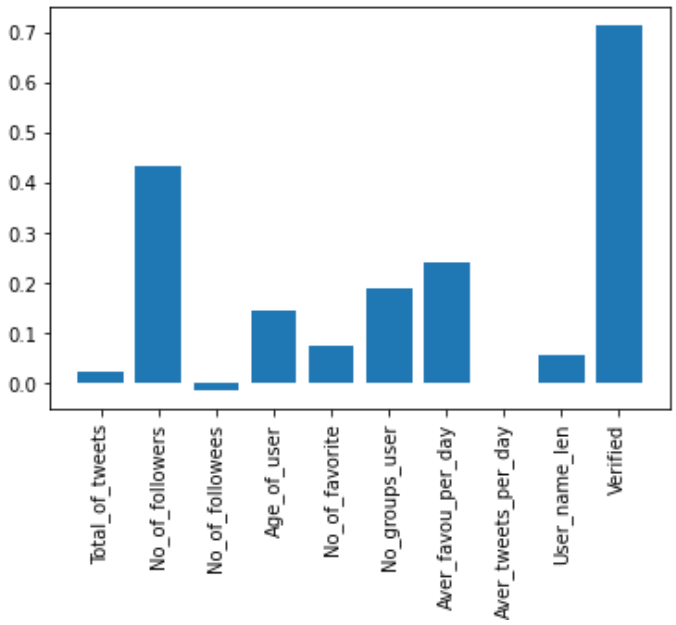

Figure 5. Result Analysis

4. CONCLUSION

In this research, it aims to build a retweet prediction system using user-based and content-based features and the Artificial Neurak Network - Genetic Algoriithm method to get pretty good results. We can see this from the results of quite good performance for split data with content-based accuracy results of 90.14%, user-based 69.11%, content-based precision value 54.20%, user-based 72.31%, content-based recall value 83.87%, user-based 56.13%, and content-based f1-score value 65.44%, user-based 59.16%. The hyperparameters tuning method that has been tested has an accuracy below the Split data scenario and does not really affect the dataset used in this study. For future research, new features such as time-based or social with classification methods can be developed.

REFERENCES

[1] H. Amarullah Purwaatmaja Ash-Shidiq EFSA and K. Muslim Lhaksmana, "Prediksi Retweet Menggunakan Fitur Berbasis Pengguna dan Fitur Berbasis Konten dengan Metode Klasifikasi ANN," vol. 8, no. 5, pp. 11207–11215, 2021.
[2] M. S. Z. S and K. M. Lhaksmana, "Prediksi Retweet Berdasarkan Feature User-Based Menggunakan Metode Klasifikasi Random Forest," 2021.
[3] J. Tugas and A. Fakultas, "Implementasi Information Gain (IG) dan Genetic Algorithm (GA)," pp. 1–16, 2021.
[4] I. M. S. Putra, "Penerapan Algoritma Genetika Dan Implementasi Dalam MATLAB," vol. 53, no. 9, pp. 1689–1699, 2018.
[5] T. B. N. Hoang and J. Mothe, "Predicting information diffusion on Twitter – Analysis of predictive features," *J. Comput. Sci.*, vol. 28, pp. 257–264, 2018, doi: 10.1016/j.jocs.2017.10.010.
[6] S. N. Firdaus, C. Ding, and A. Sadeghian, "Retweet: A popular information diffusion mechanism – A survey paper," *Online*

Copyright © 2022 **Edvan Tazul Arifin**, Page 527
BITS is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 522−528
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1931

Soc. Networks Media, vol. 6, pp. 26–40, 2018, doi: 10.1016/j.osnem.2018.04.001.
[7] A. M. A. Afinda, "Implementasi Metode Genetic Algorithm-Support Vector Machine pada Studi QSAR in-house molecules sebagai Inhibitor Papain-like Protease (PLpro) SARS-CoV-2," pp. 1–19, 2020.
[8] D. Febrianto and K. Muslim Lhaksmana, "Prediksi Retweet Dengan Fitur Berbasis Pengguna dan Tingkat Sentimen Menggunakan Metode Klasifikasi Naive Bayes," vol. 8, no. 5, pp. 11200–11206, 2021.
[9] T. Sutisna Maulasirri, Jondri, and K. Lhaksamana Muslim, "Prediksi Retweet Berbasis Konten dan Pengguna dengan Metode Klasifikasi Naïve Bayes," *e-Proceeding Eng.*, vol. 8, no. 5, pp. 11175–11181, 2021.
[10] R. A. Cahya, D. Adimanggala, and A. A. Supianto, "Deep Feature Weighting Based on Genetic Algorithm and Naïve Bayes for Twitter Sentiment Analysis," *Proc. 2019 4th Int. Conf. Sustain. Inf. Eng. Technol. SIET 2019*, pp. 326–331, 2019, doi: 10.1109/SIET48054.2019.8986107.
[11] Rakes, Jondri, and K. M. Lhaksamana, "Prediksi retweet berdasarkan feature user-based menggunakan metode klasifikasi Support Vector Machine," vol. 8, no. 5, pp. 11183–11191, 2021.
[12] I. P. Dewi, J. Jondri, and K. M. Lhaksmana, "Prediksi Retweet Menggunakan Metode Bernoulli Dan Gaussian Naive Bayes Di Media Sosial Twitter Dengan Topik Vaksinasi Covid-19," eProceedings Eng., vol. 8, no. 5, pp. 11216–11225, 2021, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/15627/15340%0Ahttps://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/15627.
[13] Idris, E. Utami, and A. D. Hartanto, "Klasifikasi Kepribadian dengan Metode DISC pada Twitter Menggunakan Algoritma Artificial Neural Network," Tecnoscienza, vol. 5, no. 1, pp. 1–20, 2020.
[14] S. Bazzaz Abkenar, E. Mahdipour, S. M. Jameii, and M. Haghi Kashani, "A hybrid classification method for Twitter spam detection based on differential evolution and random forest," Concurr. Comput. Pract. Exp., vol. 33, no. 21, pp. 1–20, 2021, doi: 10.1002/cpe.6381.
[15] R. H. Anggia 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.
[16] G. Ivan, Learning Genetic Algorithms With Python : Empower The Performance Of Machine Learning And AI Models With The Capabilities Of A Powerful Search Algorithm. 2021.
[17] I. Farkas and P. Masulli, Artificial Neural Networks and Machine Learning – ICANN 2020, vol. 12396 LNCS. 2020.

Copyright © 2022 **Edvan Tazul Arifin**, Page 528
BITS is licensed under a Creative Commons Attribution 4.0 International License