JINAV: Journal of Information and Visualization *ISSN: 2746-1440 (online)*
Vol. 4 No. 2 (2023)
https://doi.org/10.35877/454RI.jinav1903

Retweet Prediction Using Artificial Neural Network Method Optimized with Firefly Algorithm

Muhamad Rifqi Supriadi[a], Jondri[b*], Indwiarti[c]

[a]*Faculty of Informatics, Telkom University, Indonesia, mrifqisups@student.telkomuniversity.ac.id*
[b]*Faculty of Informatics, Telkom University, Indonesia, jondri@telkomuniversity.ac.id*
[c]*Faculty of Informatics, Telkom University, Indonesia, indwiarti@telkomuniversity.ac.id*

Abstract

Twitter is one of the social media platforms that has a large user base across various demographics. Users can use Twitter to search for information about celebrities, political issues, products, and trending topics of discussion. The information shared on Twitter can be referred to as tweets. Tweets can be further shared by other users using the retweet feature, which allows the tweet to reach a wider audience. This research aims to build a retweet prediction system and examine how tweets will spread. The method used in this research is Artificial Neural Network classification optimized with Firefly Algorithm, based on user-based and content-based features. This modeling approach demonstrated the best results after applying imbalanced class handling using oversampling with the SMOTE technique. The F1-Score obtained in this research is 88.07%.

Keywords: Twitter, tweet, retweet, ANN, FA, user-based, content-based.

Received: 15 July 2023 Revised: 4 October 2023 Accepted: 23 October 2023

1. Introduction

Social media has become an integral part of the daily lives of numerous internet users. These platforms serve as a primary channel for searching and sharing information, as well as communicating with people worldwide(Ramadhy & Sibaroni, 2022). Social media platforms such as Facebook and Twitter have the ability to make a topic go viral quickly(Varshney et al., 2017). Social media also facilitates rapid information retrieval and enables the exchange of opinions due to the real-time dissemination of information on these platforms. Consequently, there is a continuous flow of information available at any given time.(Rakes et al., 2021).

Twitter is one of the most widely used social media platforms in Indonesia, On Twitter, users can search for information, stay updated with the latest news, and communicate with other users. The information is conveyed through a tweet. One of the features on Twitter is retweet, which indicates a social connection between users. When a user retweets a tweet, that tweet will be shared again to their followers (Evkoski et al., 2021). As a result, the tweet will reach a wider audience as it gets shared through retweet. Modeling information diffusion is important to understand the spread of information that occurs on Twitter. (Hoang & Mothe, 2018).

Fans of K-pop are a group of people who appreciate South Korean culture, including music, dramas, and films. K-pop music has gained immense popularity internationally due to effective marketing strategies, and its music has been embraced in various parts of the world(Sarah, 2012). With the advancement of social media, K-pop fans now have numerous platforms to discuss and interact with fellow fans. These platforms provide spaces for fans to engage in conversations about their favorite idols, preferred groups, genres, dramas, and more. It has created a vibrant community where fans can connect and share their passion for K-pop. One of the social media platforms widely used by K-pop fans is Twitter. In this research, "KPOP" will be used as a keyword to collect tweet data that will be utilized for retweet prediction.

* Corresponding author.
E-mail address: jondri@telkomuniversity.ac.id

In the previous study, machine learning was applied using a multi-class classification method with three groups of features utilized in the modeling: user-based, time-based, and content-based features(Hoang & Mothe, 2018). In the user-based feature, it includes features related to the user, such as the number of followers. The time-based feature includes information about when the tweet was created. The content-based feature includes attributes related to the content of the tweet, such as whether the tweet contains videos or photos. In the study conducted by (Xu & Yang, 2012), the prediction of whether a tweet would be retweeted or not was performed using classification methods such as Decision Tree J48, Support Vector Machine, and Logistic Regression.

In this research, the author aims to predict retweets by utilizing user-based and content-based features using Artificial Neural Network (ANN) method, which will be optimized with the Firefly Algorithm.

2. Related Study

In the study conducted by (Hoang & Mothe, 2018), the research aimed to predict whether a tweet would be retweeted or not using a multi-class classification method. The study utilized three groups of features: user-based, time-based, and content-based features. In the study, the performance of the model constructed was compared to the previous research, and the results showed an improvement of 5% compared to the previous study. The most influential features identified in the study were the number of followers, the number of followees, and the number of groups followed. These features belong to the user-based feature(Hoang & Mothe, 2018).

The next research, titled "*Analyzing User Retweet Behavior on Twitter"*, the research focuses on analyzing behavior of users who engage in retweet. This study aims to analyze whether a tweet will be retweeted by specific users using several methods for comparison, namely Decision Tree J48, Support Vector Machine, and Logistic Regression. The study utilizes four feature groups: social-based, content-based, tweet-based, and context-based. The researchers used the leave-one-feature-out comparison to identify the influential factors in user behavior when retweeting. The results showed that user-based features were significant in predicting retweets(Xu & Yang, 2012).

In 2022, a research study titled "*Prediction Retweet Using User-Based and Content-Based with ANN-GA Classification Method*" was conducted to predict retweets using user-based and content-based features. The method employed in this research was Artificial Neural Network (ANN) optimized with Genetic Algorithm (GA). In this study, separate modelling was conducted for user-based and content-based features to determine which feature set performed better. The results of the research showed that content-based features outperformed user-based features, achieving an F1-Score of 65.44%, while user-based features achieved a score of 59.16%.

2.1. Twitter

Twitter was founded by Jack Dorsey, Evan Williams, Noah Glass, and Biz Stone on March 21, 2006. Twitter is a social media platform that allows users to follow other users and receive the information they share. The information shared on Twitter is referred to as tweets. A tweet can contain more than 140 characters and can include up to 4 photos or 1 video. Tweets can also be retweeted by other users, allowing them to share the tweet with their followers and thereby reaching a wider audience.

2.2. Features

The features that will be used in the study are the most influential features on retweets, namely user-based and content-based features. These features are the most impactful for tweets to receive retweets. (Jenders et al., 2013).

User-based features are related to the user, and here are the features included in the user-based category:

- Number of Followers
- Number of Followees
- Number of Tweets
- Account Age
- Number of Characters in Username
- Number of Characters in Bio
- Account Verification Status

– Average Number of Tweets per Day

Content-based features are related to the content of the tweet. Here are the features included in the content-based category:

– Tweet Length: Number of characters in the tweet.
– Media: Whether the tweet contains photos or video.
– Hashtags: Whether the tweet includes hashtags.
– Mentions: Whether the tweet mentions other users.
– Number of Likes: Number of users who liked the tweet.
– Capitalization: Whether the tweet starts with a capital letter.
– Numbers: Whether the tweet contains numbers.
– Exclamation Marks: Whether the tweet includes exclamation marks.
– Retweet Suggest: Whether the tweet includes a suggestion to retweet.

2.3. Artificial Neural Network (ANN)

ANN is a mathematical model that simulates the structure and function of biological neural networks. (Suzuki, 2011). ANN consists of neurons that are interconnected to form a network. Each neuron in the network is connected to other neurons, forming a complex network structure. (Simon Haykin (McMaster University, Hamilton, Ontario, 2005). In the study conducted by (Sonali & Wankar, 2014) it is mentioned that ANN has several advantages, including:

– Adaptive Learning

  ANN can learn how to solve problems based on the given training data or initial experience. It has the ability to adjust its learning based on the input patterns and make necessary changes in its weights and connections.

– Self-Organisation

  ANN can create representations of the information received during the learning process. It can organize and structure the learned information in a meaningful way.

Furthermore, ANN typically comprises three layers as depicted in Figure 1. The first layer is the Input Layer, where the network receives the input data. The next layer is the Hidden Layer, which performs computations and transformations on the input data. Finally, there is the Output Layer, which produces the final output or prediction based on the computations performed in the Hidden Layer(Nielsen, 2015).

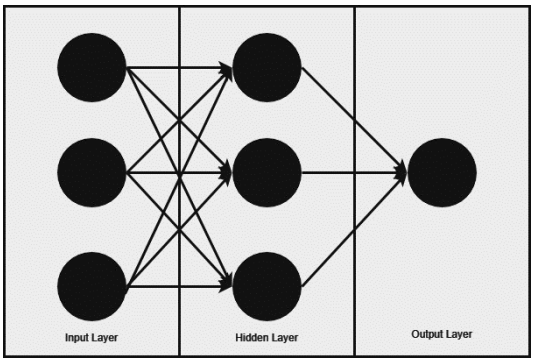

Fig. 1. Diagram Artificial Neural Network (ANN)

– Input Layer

  This layer functions as the starting point of a network, where the neurons in this layer receive data and pass it on to the next layer.

– Hidden Layer

  This layer is located between the input and output layers, and it performs the process of combining data from the previous layer. Each hidden layer consists of interconnected neurons.

- Output Layer

  The neurons in this layer generate predictions based on the data received from the previous layers.

2.4. Firefly Algorithm (FA)

The Firefly Algorithm (FA) is inspired by the communication and social behavior of fireflies. Fireflies emit light from their tails, and those with higher light intensity attract fireflies with lower intensity. The light emitted by fireflies serves two purposes: to attract the attention of other fireflies as potential mates and to trap prey(Yang, 2010). The Firefly Algorithm follows three main rules:(Pan et al., 2014):

- Fireflies are unisex, meaning they can be attracted to one another without considering gender.
- The attractiveness of fireflies is proportional to the intensity of their light. Fireflies with lower light intensity are attracted to those with higher intensity. If a firefly cannot find another firefly with higher intensity, it will move randomly.
- The light intensity of fireflies is determined by an objective function.

Here is the formula for attractiveness in the Firefly Algorithm (FA):

$$\beta(r) = \beta_0 e^{-\gamma r^2} \quad (1)$$

In this case, the attractiveness of fireflies at distance r, denoted as β(r), is defined. β_0 represents the attractiveness of fireflies at r=0, γ is the light absorption coefficient, and r represents the distance between the source firefly and the observing firefly. In this research, the Firefly Algorithm (FA) is utilized for hyperparameter tuning of the Artificial Neural Network.

2.5. Confusion Matrix

The Confusion Matrix is a technique used to summarize the performance of a classification algorithm. (Brownlee, 2016). The Confusion Matrix has several terms that will be used in the matrix:

- True Positive (TP)

  The condition where the model predicts positive and the result is correct. In this research, it means the model predicts that a tweet will be retweeted, and indeed the tweet does receive retweets.

- False Positive (FP)

  The condition where the model predicts positive, but it should have been negative. In this research, it means that the model predicts that a tweet will receive retweets, but in reality, the tweet does not receive any retweets.

- True Negative (TN)

  The condition where the model predicts negative and the result is correct. In this research, it means the model predicts that a tweet will not receive retweets, and indeed the tweet does not receive any retweets.

- False Negative (FN)

  The condition where the model predicts negative, but it should have been positive. In this research, it means the model predicts that a tweet will not have any retweets, but in reality, the tweet does have retweets.

The matrix obtained using the Confusion Matrix includes metrics such as Accuracy, Precision, Recall, and F-Measure/F1-Score. Here are the formulas for these metrics:

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

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

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (4)$$

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

3. Built System

In this stage, the system to be built will be explained, and the figure 2 is the flowchart of the system to be constructed.

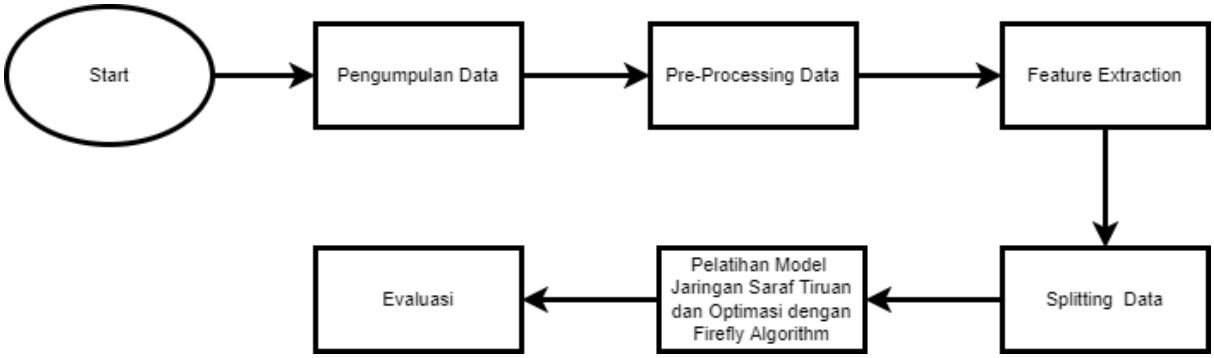

Fig. 2. Flowchart System.

3.1. Data Crawling

Data crawling was conducted through the website Netlytics.org. The collected dataset consists of 7135 tweets in the Indonesian language with the search keyword "KPOP" and covers the time span of October to November 2022.

3.2. Pre-processing Data

- Data filtering was performed to exclude unnecessary tweet types such as "retweet" and "quote" from the combined dataset. Only the "original" tweet type was selected as required.
- Labeling the data involves creating additional features that are not present in the dataset, one of which is the "retweeted" feature to indicate whether a tweet has been retweeted or not. Class 0 represents tweets that do not have any retweets, while class 1 represents tweets that have received retweets.

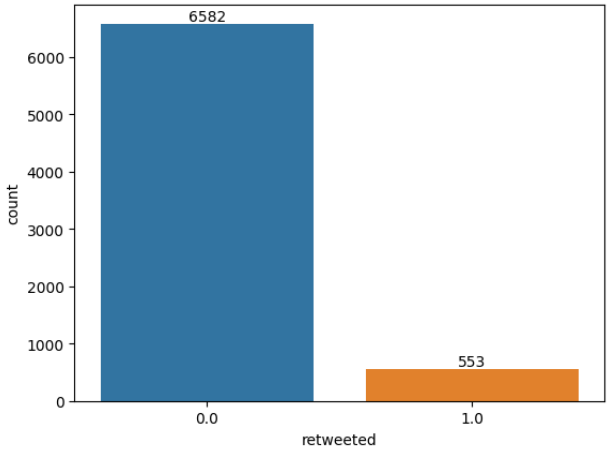

Fig. 3. Retweeted Class Distribution

- Data normalization will be performed using StandardScaler to expedite the classification process.

3.3. Feature Extraction

After the pre-processing stage, feature extraction is performed to select important features that will be used for classification. The features that will be used are:

Table 1. Feature Extraction

Feature	Description	Data Type
total_tweet	Total number of tweets made by the user	Numeric
followers_count	Number of people following the user	Numeric
followee_count	Number of people the user is following	Numeric
avg_tweet_per_day	Average number of tweets per day	Numeric
age_of_account	Number of days since the user created the account	Numeric
uname_length	Character length of the username	Numeric
bio_length	Character length of the user's bio	Numeric
verified	Indicates whether the user account is verified or not	Boolean
favorite_count	Number of likes on the tweet	Numeric
tweet_length	Character length of the tweet	Numeric
opt_length	Indicates whether the tweet length is between 70-100 characters	Boolean
contain_upper	Indicates whether the tweet contains uppercase letters	Boolean
contain_user_mentioned	Indicates whether the tweet contains user mentions	Boolean
contain_hashtag	Indicates whether the tweet contains hashtags	Boolean
contain_number	Indicates whether the tweet contains numbers	Boolean
contain_excl	Indicates whether the tweet contains exclamation marks	Boolean
contain_rt_suggest	Indicates whether the tweet contains suggestions for retweeting	Boolean
retweeted	Label 0 for tweets without retweets and 1 for tweets with retweets	Boolean

3.4. Hyperparameter Tuning

Hyperparameter tuning is performed to obtain the best combination of hyperparameters for the ANN model, aiming to achieve good performance results(Bardenet et al., 2013), In this study, hyperparameter tuning is performed using MLPClassifier and optimized with Firefly Algorithm. Here is the table of hyperparameters used for this research:

Table 2. Hyperparameter

Hyperparameter	Value
Hidden Layer Size	$y_i = \lfloor 10 + x_i * 40 \rfloor \in i[1,4], y_i[10,50]$
Activation	$y_5 = \lfloor x_4 \rfloor \in y_5[0,1]$
Solver	$y_6 = \lfloor x_5 \rfloor \in y_6[0,1]$
Alpha	$y_7 = x_6 * 0.0001$
Learning rate initial	$y_8 = 0.001 + x_7$

In the hidden layer, there will be 4 values obtained, and these values will be between the range of 10 to 50. For the activation parameter, if the value is 0, the activation function will be 'relu'. If the value is 1, the activation function will be 'tanh'. For the solver parameter, if the value is 0, the solver will be 'adam'. If the value is 1, the solver will be 'lbfgs'. The value of alpha is calculated as $x_6 * 0.0001$ and the initial learning rate is $0.001 + x_7$.

3.5. Scenario Test

– *1st Scenario Test*

In scenario 1, the testing is conducted on a dataset that has not undergone imbalanced handling.

– *2nd Scenario Test*

In 2nd scenario, the testing will be conducted using the undersampling method using RandomUnderSampler. Undersampling is performed to balance the classes by randomly eliminating data from the majority class (Kotsiantis et al., 2006). In this dataset, the class that receives retweets is 553.

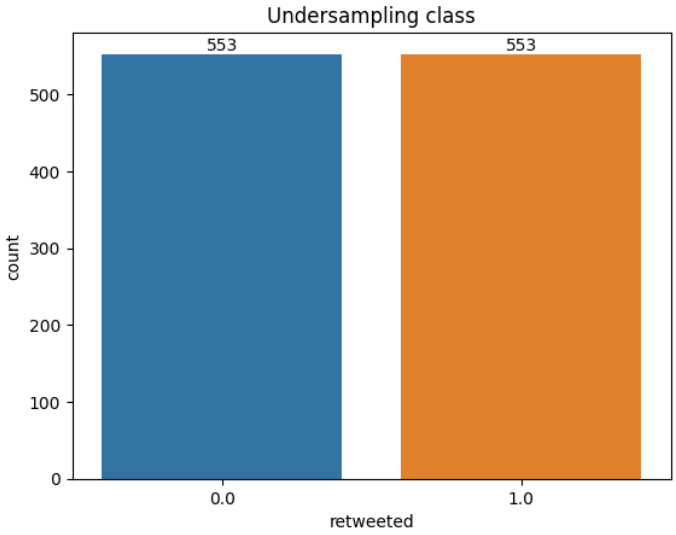

Fig. 4. Undersampling Class

– *3rd Scenario Test*

In 3rd Scenario, testing is conducted using the oversampling method using the SMOTE (Synthetic Minority Oversampling Technique) technique. Oversampling with SMOTE is performed to generate synthetic data samples that will be added to the minority class.(Fernández et al., 2018). In this dataset, the class that does not receive retweets is 6582.

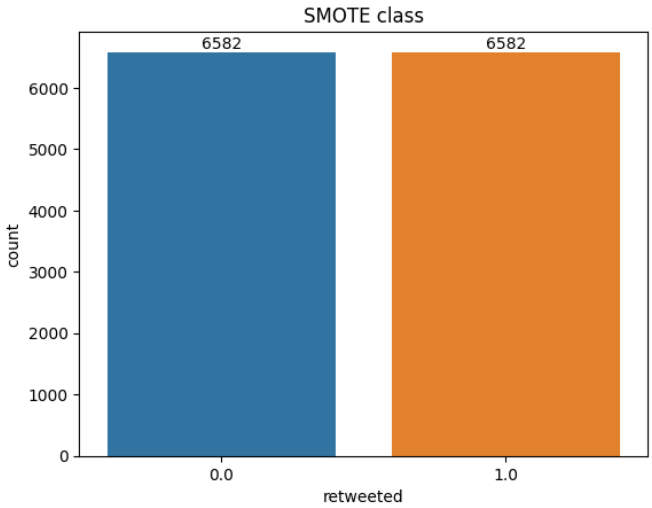

Fig. 5. SMOTE Class

4. Test Result & Evaluation

– *1st Scenario*

In the first test, using the dataset that has not undergone imbalanced handling, the hyperparameter tuning conducted in this test yielded the results as shown in Table 3.

Table 3. Imbalanced class best hyperparameter

Class	Hidden Layer Size	Activation	Solver	Alpha	Learning rate initial
Imbalanced Class	(15, 10, 41, 50)	tanh	Adam	0	0.002359349135353122

Table 4. Imbalanced test result

Class	Accuracy	Precision	Recall	F1-Score
Imbalanced Class	92.85%	68%	15.32%	25%

The model built with these hyperparameters achieved a high accuracy of 92.85%. The precision value obtained is quite good, with a value of 68%. However, the recall value is low at 15.32% due to the high number of False Negatives in the confusion matrix. As a result, the F1-score obtained is also low at 25%.

– 2nd Scenario

In the second test, the dataset was handled using RandomUnderSampler, and the hyperparameter tuning conducted with this dataset yielded the results as shown in Table 5.

Table 5. Undersampling class best hyperparameter

Class	Hidden Layer Size	Activation	Solver	Alpha	Learning rate initial
Undersampling Class	(43, 11, 29, 41)	tanh	Adam	2.55	0.0012857531320076452

Table 6. Undersampling test result

Class	Accuracy	Precision	Recall	F1-Score
Undersampling Class	76.87%	18.26%	56.76%	27.63%

The model built with these hyperparameters shows subpar performance with a decrease in accuracy and precision. The low precision value in this test is due to the increase in false positives after undersampling, which negatively impacts precision. However, the recall value improves significantly to 56.76% as the balanced dataset reduces false negatives. The F1-score shows a slight improvement with a value of 27.63%. Based on these performance results, it can be concluded that undersampling in this study is still not able to achieve satisfactory performance, with only recall showing significant improvement.

– 3rd Scenario

In the third test, the dataset used underwent oversampling using the SMOTE technique. The hyperparameter tuning conducted with this dataset yielded the parameters as shown in Table 7.

Table 7. SMOTE class best hyperparameter

Class	Hidden Layer Size	Activation	Solver	Alpha	Learning rate initial
SMOTE Class	(39, 36, 48, 41)	tanh	Adam	2.13	0.00298107063156789

Table 8. SMOTE test result

Class	Accuracy	Precision	Recall	F1-Score
SMOTE Class	97.97%	81.06%	96.4%	88.07%

The model built with these hyperparameters demonstrates excellent performance, as indicated by improvements in all metrics. The accuracy value in this scenario shows an increase compared to the accuracy value in scenario 1. The precision value in this model also shows excellent results, indicating that oversampling with SMOTE can reduce the

false positive rate. Similar to scenario 2, the balanced data reduces the false negative rate, resulting in improved recall in this scenario. With good precision and recall values, the F1-Score increases to 88.07%.

5. Conclusion

In this study, it can be concluded that the Artificial Neural Network (ANN) optimized with the Firefly Algorithm yielded good results, with the best outcome observed in scenario 3, where the test was conducted using a dataset that underwent oversampling with the SMOTE technique. Through oversampling, the model achieved an accuracy of 97.97%, precision of 81.06%, recall of 96.4%, and an F1-Score of 88.07%. For future research, it is recommended to explore alternative optimization algorithms for the Artificial Neural Network.

References

Bardenet, R., Brendel, M., Kégl, B., & Sebag, M. (2013). Collaborative hyperparameter tuning. *30th International Conference on Machine Learning, ICML 2013*, *28*(PART 2), 858–866.

Brownlee, J. (2016). *What is a Confusion Matrix in Machine Learning?* https://machinelearningmastery.com/confusion-matrix-machine-learning/

Evkoski, B., Mozetič, I., Ljubešić, N., & Novak, P. K. (2021). Community evolution in retweet networks. *PLoS ONE*, *16*(9 September), 1–21. https://doi.org/10.1371/journal.pone.0256175

Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, *61*, 863–905. https://doi.org/10.1613/jair.1.11192

Hoang, T. B. N., & Mothe, J. (2018). Predicting information diffusion on Twitter – Analysis of predictive features. *Journal of Computational Science*, *28*, 257–264. https://doi.org/10.1016/j.jocs.2017.10.010

Jenders, M., Kasneci, G., & Naumann, F. (2013). Analyzing and predicting viral tweets. *WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web*, 657–664. https://doi.org/10.1145/2487788.2488017

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets : A review. *Science*, *30*(1), 25–36. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.96.9248&rep=rep1&type=pdf

Nielsen, M. (2015). Neural Networks and Deep Learning. In *Determination press San Francisco, CA, USA*. Determination press San Francisco, CA, USA. https://doi.org/10.1108/978-1-83909-694-520211010

Pan, Q., Darabos, C., Moore, J., & Yang, X. (2014). Cuckoo Search and Firefly Algorithm Theory and Applications. In *Studies in Computational Intelligence* (Vol. 516). http://dx.doi.org/10.1007/978-3-642-29066-4_11

Rakes, Jondri, & Lhaksmana, K. M. (2021). Prediksi Retweet Berdasarkan Feature User-based Menggunakan Metode Klasifikasi Random Forest. *EProceedings …*, *8*(5), 11192–11199. https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/15633%0Ahttps://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/15633/15346

Ramadhy, I. F., & Sibaroni, Y. (2022). Analisis Trending Topik Twitter dengan Fitur Ekspansi FastText Menggunakan Metode Logistic Regression. *JURIKOM (Jurnal Riset Komputer)*, *9*(1), 1. https://doi.org/10.30865/jurikom.v9i1.3791

Sarah, L. (2012). Catching the K-Pop Wave: Globality in the Production, Distribution and Consumption of South Korean Popular Music. *Senior Capstone Projects*, 149.

Simon Haykin (McMaster University, Hamilton, Ontario, C. (2005). *Neural Networks - A Comprehensive Foundation - Simon Haykin.pdf* (p. 823).

Sonali, B. M., & Wankar, P. (2014). Research Paper on Basic of Artificial Neural Network. *International Journal on Recent and Innovation Trends in Computing and Communication*, *2*(1), 96–100.

Suzuki, K. (2011). *Artificial Neural Networks : Methodological Advances and Biomedical Applications*. IntechOpen. https://doi.org/10.5772/644

Varshney, D., Kumar, S., & Gupta, V. (2017). Predicting information diffusion probabilities in social networks: A Bayesian networks based approach. *Knowledge-Based Systems*, *133*, 66–76. https://doi.org/10.1016/j.knosys.2017.07.003

Xu, Z., & Yang, Q. (2012). Analyzing user retweet behavior on twitter. *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012*, 46–50. https://doi.org/10.1109/ASONAM.2012.18

Yang, X. S. (2010). Firefly algorithm, Lévy flights and global optimization. *Research and Development in Intelligent Systems XXVI: Incorporating Applications and Innovations in Intelligent Systems XVII*, 209–218. https://doi.org/10.1007/978-1-84882-983-1_15