Prediction Retweet Using User-Based, Content-Based and Time-Based Features with ANN-Firefly Classification Method

Alif Aqshal[1*], Indri Anugrah Ramadhani[2]
[1-2] Faculty of Exact Sciences Education, Universitas Pendidikan Muhammadiyah Sorong, Papua, Indonesia

Corresponding Author's e-mail : *alifaqshal@unimudasorong.ac.id*

ARMADA
JURNAL PENELITIAN MULTIDISIPLIN

e-ISSN: 2964-2981
ARMADA : Jurnal Penelitian Multidisiplin
https://ejournal.45mataram.ac.id/index.php/armada
Vol. 04, No. 05 Mei, 2026
Page: 426-432

DOI:
https://doi.org/10.55681/armada.v4i5.2014

Article History:
Received: Maret 01, 2026
Revised: April 08, 2026
Accepted: Mei 19, 2026

Abstrak: *Retweets on the Twitter platform play an important role in the dissemination of information. However, there are still limitations in effective prediction systems to estimate whether a tweet will be retweeted or not. This study aims to develop a retweet prediction model by utilizing a combination of user-based, content-based, and time-based features. The model was built using an Artificial Neural Network (ANN) and optimized using the Firefly Algorithm (FA) to improve classification accuracy. The dataset was collected from Twitter using data crawling techniques through the Tweepy library with a focus on the keyword "Covid Vaccination", resulting in 12,796 Indonesian-language tweets. The obtained data then went through a preprocessing stage, including text cleaning, tokenization, normalization, and stemming. Next, an ANN-FA model was trained to classify the likelihood of a tweet being retweeted. The experimental results showed that the ANN-FA model achieved an accuracy rate of 90.29%, which was higher than the baseline ANN model without optimization. These findings indicate that the application of the Firefly Algorithm can significantly improve classification performance. The contribution of this research lies in the development of a retweet prediction system that integrates multidimensional features with metaheuristic optimization, which can be utilized to support digital information dissemination strategies on social media platforms.*

Kata Kunci: *Retweet Prediction, Artificial Neural Network, Firefly Algorithm, Twitter*

INTRODUCTION

The development of social media has accelerated the dissemination of information, including information related to COVID-19. As of January 2022, the number of active social media users in Indonesia reached 191 million, representing an increase of 12.35% compared to the previous year (Febiansyah et al., 2025). Social media functions as a digital platform that facilitates self-expression, talent development, and the dissemination of information. The ability to share content in real time is a significant feature that enhances information flow. This technological convenience can also be exploited by certain entities to serve specific interests, enabling the formation of public opinion through targeted narratives (Arifin et al., 2022). One of Twitter's key features for information dissemination is the retweet function, where users create posts known as tweets that can be reposted by other users. Retweets enable content to spread widely, often driven by user interest or approval, as reflected in likes and further shares (Anggia & Muslim, 2021).

The number of tweets shared on Twitter is remarkably high, with some receiving significant retweet engagement while others receive little to none. This phenomenon can serve as an important subject of academic research in exploring online communication, user engagement, and information dissemination patterns. Such studies are particularly relevant in areas such as

marketing, business analysis, and predictive decision-making, where user interaction data provide valuable insights (Akbar, 2023). Twitter is also widely used among researchers and developers because of its accessibility and the availability of relevant public data. Information on this platform is disseminated through the retweet feature, allowing content to spread rapidly across user networks. The greater the number of retweets a tweet receives, the wider its reach and influence become (Syah Zannuar & Lhaksmana, 2021).

In 2022, Julizar Wiranto Harahap et al. conducted a study entitled *Microarray-Based Classification Model for Identifying Parkinson's Disease Using the Firefly Algorithm–Support Vector Machine Method*. The objective of the study was to classify the relationship between blood expression and indications of Parkinson's disease. The findings demonstrated that the Firefly Algorithm produced the highest accuracy, achieving a performance rate of 68.57% (Harahap et al., 2022). In the same year, Edvan Tazul Arifin et al. conducted research entitled *Retweet Prediction Based on User-Based and Content-Based Features Using the ANN-GA Method*. The study aimed to predict retweet behavior and achieved an accuracy rate of 90% (Arifin et al., 2022). Furthermore, Muhammad Syah Zannuar S. and K. M. Lhaksmana (2021) developed a retweet prediction system using user-based features and the Random Forest classification method, which achieved an accuracy level of 70%.

Earlier research conducted by Hoang and Mothe (2018), entitled *Predicting Information Diffusion on Twitter – Analysis of Predictive Features*, aimed to predict whether a tweet would spread further and measure its diffusion level across the platform. The prediction model incorporated user-based, time-based, and content-based features, while the classification process employed the Random Forest algorithm. The experimental results demonstrated highly satisfactory predictive performance. Based on these previous studies, the present research employs the Artificial Neural Network (ANN) classification method optimized using the Firefly Algorithm. ANN is recognized for its strong predictive capability, ability to model complex relationships, and robustness against noisy data (Arifin et al., 2022). In this study, the Firefly Algorithm is utilized to optimize ANN weights because of its effective global exploration capability and efficiency in identifying optimal solutions within large search spaces (Kumar & Kumar, 2020).

RESEARCH METHODS

This study proposes a retweet prediction model using Artificial Neural Network (ANN) optimized with the Firefly Algorithm (FA). The research process consists of data collection, preprocessing, feature extraction, classification, optimization, and evaluation stages, as illustrated in Figure 1. Data were collected from Twitter through the official Twitter API using the Tweepy library with the keyword *"Vaksinasi Covid"*, resulting in a dataset of 12,796 tweets. The collected data were subsequently preprocessed through case folding, data cleaning, tokenization, normalization, and stemming to improve data consistency and reduce noise. In addition, missing value handling, duplicate removal, outlier detection, and class imbalance checking were conducted to ensure dataset quality before the modeling stage. After preprocessing, the dataset was divided into training and testing data to evaluate model generalization and predictive performance. This study employed three categories of retweet features, namely user-based, content-based, and time-based features. User-based features describe account characteristics and user activity, content-based features represent tweet properties and sentiment information, while time-based features provide contextual information regarding posting time. The complete list of features used in this study is presented in Table 1.

The classification process utilized ANN as the primary predictive model because of its capability to capture complex and nonlinear relationships within social media data. To improve model performance, the Firefly Algorithm was integrated into the ANN training process to optimize model parameters and identify better weight configurations iteratively. The movement of fireflies during optimization was determined based on attractiveness and fitness values, enabling the algorithm to converge toward optimal solutions. Model performance was evaluated using a confusion matrix consisting of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Several evaluation metrics, including accuracy, precision, recall, and F1-score, were employed to assess classification performance comprehensively. Through this approach, the

proposed ANN-FA framework is expected to provide robust and accurate retweet prediction performance in analyzing information dissemination patterns on Twitter.

RESULT AND DISCUSSION

This study utilized a dataset consisting of 12,796 tweets collected from Twitter, incorporating three categories of features, namely user-based, content-based, and time-based features, to capture various aspects influencing retweet behavior. User-based features describe user activity and account characteristics, content-based features represent tweet properties and sentiment information, while time-based features provide contextual information related to posting time. To evaluate the effectiveness and robustness of the proposed Artificial Neural Network–Firefly Algorithm (ANN-FA) model, two evaluation scenarios were implemented, namely dataset splitting and K-Fold Cross Validation. These evaluation strategies were employed to assess model stability, generalization capability, and predictive consistency across different data partitioning schemes.

Test Result and Analysis of Scenario 1

At the initial stage, three data splitting scenarios were applied to evaluate the performance of the proposed model. The dataset was divided into training and testing sets using ratios of 70:30, 80:20, and 90:10. In each scenario, the Firefly Algorithm was used to optimize the model parameters. To ensure consistent results, each configuration was tested 10 times. The evaluation focused on accuracy and F1-score to measure how well the model classified retweet behavior using user-based, content-based, and time-based features.

Table 1. Split Dataset

Split Dataset	Accuracy	F1-Score
70:30	99,7%	99,5%
80:20	99,6%	99,4%
90:10	99,7%	99,6%

Based on Table 3, the proposed Artificial Neural Network–Firefly Algorithm (ANN-FA) model demonstrated consistently high performance across all dataset splitting scenarios, with accuracy and F1-score values exceeding 99%. The 70:30 split configuration achieved an accuracy of 99.7% and an F1-score of 99.5%, while the 80:20 split showed only a slight decrease in performance. Meanwhile, the 90:10 split again produced the highest performance values among all scenarios. Although the 90:10 configuration yielded the best results, the differences among the three scenarios were relatively marginal, indicating that the proposed model maintained stable predictive capability regardless of the data partitioning strategy. This finding suggests that the model possesses strong generalization ability and does not rely excessively on a particular training–testing composition. In machine learning research, stable performance across multiple validation scenarios is considered an important indicator of model robustness and reliability, particularly in classification tasks involving social media data characterized by high variability and noise (Kumar & Kumar, 2020).

The strong performance achieved in this study is likely influenced by the comprehensive feature engineering strategy employed. The integration of user-based, content-based, and time-based features enabled the model to capture multidimensional characteristics associated with retweet behavior. User-based features represent user popularity, activity, and influence within the social network, while content-based features describe tweet characteristics such as sentiment, tweet structure, and contextual relevance. Meanwhile, time-based features provide temporal information regarding posting behavior and audience activity patterns. Previous studies have emphasized that combining heterogeneous feature categories can significantly improve retweet prediction performance because retweet behavior is influenced not only by content quality but also by user reputation and posting context (Hoang & Mothe, 2018). Therefore, the use of multidimensional features in this study contributed substantially to improving the discriminative capability of the model.

In addition to feature engineering, the optimization process using the Firefly Algorithm also played a significant role in enhancing model performance. ANN is widely recognized for its ability to model nonlinear and complex relationships within large-scale datasets; however, its predictive performance is highly dependent on parameter optimization and weight initialization. The Firefly Algorithm improves this process by exploring the search space iteratively and identifying optimal parameter configurations during training. Metaheuristic optimization methods such as FA have been reported to improve convergence quality, reduce local optimum trapping, and enhance predictive stability in ANN-based classification systems (Alhaidar & Aldrajy, 2024). Consequently, the integration of ANN and FA in this study enabled the model to achieve highly consistent performance across different validation scenarios.

Nevertheless, the extremely high accuracy values obtained in this study should be interpreted cautiously. Performance values approaching 100% may indicate the possibility of overfitting, where the model learns dataset-specific patterns too closely and may experience reduced performance when applied to unseen or more heterogeneous datasets. This issue is common in machine learning models trained on relatively homogeneous or topic-specific social media datasets (Lopez et al., 2022). Therefore, further evaluation using larger, more diverse, and cross-domain datasets is necessary to confirm the external validity and generalizability of the proposed model. Additional validation using independent datasets and real-time Twitter streams may also provide a more comprehensive assessment of the model's robustness under different information dissemination conditions.

Overall, the findings demonstrate that the proposed ANN-FA framework provides strong and stable retweet prediction performance across multiple dataset splitting scenarios. The combination of multidimensional features and metaheuristic optimization successfully improved classification effectiveness and model consistency. These results indicate that the proposed approach has considerable potential for supporting social media analytics, information diffusion prediction, digital marketing strategies, and online public opinion monitoring in large-scale social network environments.

The Result and Analysis of Scenario 2

In the second scenario, the model was evaluated using K-Fold Cross Validation with two different values, namely K = 5 and K = 10. This approach was used to further examine the stability of the model by testing it across multiple data partitions. The Firefly Algorithm was applied in each scenario to optimize the model parameters during the training process. Each configuration was carried out according to the number of folds, allowing every part of the dataset to be used alternately as training and testing data.

Table 2. Cross-Fold Validation

K-Fold Val.	Acc.	F1-Score
K = 5	99,7%	99,66%
K = 10	99,7%	99,6%

Based on Table 4, both K = 5 and K = 10 cross-validation scenarios produced identical performance results, with an accuracy of 99.7% and an F1-score of 99.6%. These findings indicate that the proposed Artificial Neural Network–Firefly Algorithm (ANN-FA) model demonstrates a high level of stability across different cross-validation configurations. The absence of significant performance variation between the two validation settings suggests that the model is capable of maintaining consistent predictive capability regardless of how the dataset is partitioned. In machine learning evaluation, consistent results across multiple cross-validation schemes are commonly interpreted as evidence of model robustness and reliable generalization performance, particularly in classification problems involving high-dimensional social media data (Mepaiyeda & Oluwayumi, 2023). The results further imply that both K = 5 and K = 10 configurations adequately represent the overall data distribution, enabling the model to learn stable decision boundaries and avoid excessive sensitivity to sampling variation.

The stability observed in this study is strongly associated with the integration of ANN and the Firefly Algorithm during the optimization process. ANN possesses strong capability in learning nonlinear relationships and extracting hidden patterns from complex datasets; however, its performance is often influenced by parameter initialization and convergence quality. The Firefly Algorithm contributes by optimizing network parameters iteratively through a metaheuristic search mechanism, enabling the model to identify more optimal solutions and reduce the risk of poor local minima convergence. Previous studies have shown that hybrid ANN–metaheuristic approaches can improve classification stability, accelerate convergence, and enhance predictive consistency across multiple evaluation scenarios (Supandi et al., 2023). Therefore, the highly similar results obtained under different K-Fold settings indicate that the optimization process successfully improved the reliability of the ANN model and reduced dependency on specific validation configurations.

Another important finding is that the proposed model demonstrates low sensitivity toward variations in data partitioning. In predictive modeling, highly sensitive models tend to produce fluctuating performance when exposed to different validation folds, indicating instability in learning patterns from the dataset. Conversely, the ANN-FA framework in this study maintained nearly identical performance metrics across all folds, suggesting that the extracted retweet features consistently contribute meaningful information for classification. The integration of user-based, content-based, and time-based features likely strengthened the model's ability to capture diverse behavioral patterns associated with retweet activities. This finding aligns with previous research emphasizing that multidimensional feature representation can improve predictive robustness in social media analytics and information diffusion modeling (Khoerunnisa & Astuti, 2021).

Nevertheless, despite the strong performance obtained, the extremely high evaluation scores should still be interpreted cautiously. Accuracy values approaching 100% may indicate that the dataset possesses relatively homogeneous patterns or that the model has learned highly specific characteristics of the training data. Although cross-validation reduces the likelihood of overfitting compared with a single split evaluation, further experiments using more heterogeneous datasets and cross-topic validation are still necessary to evaluate the broader generalization capability of the proposed model. Additional testing on multilingual datasets, real-time Twitter streams, or different social media platforms could provide a more comprehensive understanding of the model's robustness in practical applications. Overall, the K-Fold Cross Validation results reinforce the findings from the split dataset experiments, confirming that the proposed ANN-FA framework achieves stable, reliable, and highly consistent retweet prediction performance across different evaluation strategies.

CONCLUSION

The findings demonstrate that the proposed ANN–Firefly Algorithm framework provides highly robust and consistent performance for retweet prediction across multiple evaluation scenarios, achieving accuracy and F1-score values exceeding 99%. The integration of user-based, content-based, and time-based features effectively captures multidimensional retweet behavior patterns, while the Firefly Algorithm enhances ANN optimization and predictive stability. These results confirm the effectiveness of hybrid metaheuristic-based neural network approaches in improving social media prediction tasks and information diffusion modelling.

Despite the promising results, further evaluation using larger, more heterogeneous, and cross-domain datasets is necessary to validate the model's generalizability and reduce potential overfitting risks. Future studies are encouraged to integrate deep learning architectures, contextual interaction features, and real-time social media streams to improve scalability and practical applicability in dynamic online environments.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to all parties who contributed to this research, particularly to those who supported the data collection, analysis process, and completion

of this study. Appreciation is also extended to the academic institutions and researchers whose insights and previous studies provided valuable foundations for this research.

REFERENCES

Arifin, E. T., Jondri, & Indwiarti. (2022). *Prediksi retweet menggunakan fitur user based dan content based dengan metode klasifikasi ANN-GA*.

Alhaidar, Q. J., & Aldrajy, S. M. (2024). Neural network based phishing website detection with using Firefly Algorithm. *International Journal of Intelligent Systems and Applications in Engineering, 12*(2), 145–154.

Chao, C. F., & Horng, M. H. (2015). The construction of support vector machine classifier using the Firefly Algorithm. *Computational Intelligence and Neuroscience, 2015*, 1–10. https://doi.org/10.1155/2015/147640

Febiansyah, M., Jondri, & Indwiarti. (2025). *Prediksi retweet berdasarkan konten dan berbasis pengguna dengan metode seleksi classifier*.

Firdaus, S. N., Ding, C., & Sadeghian, A. (2018). Retweet: A popular information diffusion mechanism – A survey paper. *Online Social Networks and Media, 6*, 26–40. https://doi.org/10.1016/j.osnem.2018.04.001

Harahap, J. W., Kurniawan, I., & Nitha, F. (2022). *Model klasifikasi berbasis microarray pada identifikasi Parkinson dengan menggunakan metode Firefly Algorithm-Support Vector Machine*.

Hoang, T. B. N., & Mothe, J. (2018). Predicting information diffusion on Twitter: Analysis of predictive features. *Journal of Computer Science, 28*, 257–264. https://doi.org/10.1016/j.jocs.2017.11.002

Khoerunnisa, G., & Astuti, W. (2021). Prediction of retweets based on user, content, and time features using EUSBoost. *Journal of Physics: Conference Series, 1845*(1), 012028. https://doi.org/10.1088/1742-6596/1845/1/012028

Kumar, V., & Kumar, D. (2020). A systematic review on Firefly Algorithm: Past, present, and future. *Archives of Computational Methods in Engineering, 27*(2), 547–569. https://doi.org/10.1007/s11831-018-9303-1

Lopez, O. A. M., Lopez, A. M., & Crossa, J. (2022). Multivariate statistical machine learning methods for genomic prediction. *Frontiers in Genetics, 13*, 867780. https://doi.org/10.3389/fgene.2022.867780

Mepaiyeda, E. B., & Oluwayumi, I. A. (2023). Prediction of gas hydrate formation temperature in pipelines using artificial neural network (ANN) and Firefly Algorithm (FA). *Results in Engineering, 19*, 101302. https://doi.org/10.1016/j.rineng.2023.101302

Muhalani, R., Jondri, & Indwiarti. (2025). *Prediksi retweet berdasarkan fitur user-based, content-based, dan time-based menggunakan metode ANN-GSO*.

Romzi, M. N., & Atastina, I. (2022). *Analisis sentiment judul berita ekonomi terhadap indeks harga saham gabungan menggunakan metode Long Short-Term Memory*.

Supandi, M. R., Jondri, & Indwiarti. (2023). Retweet prediction using artificial neural network method optimized with Firefly Algorithm. *International Journal of Intelligent Engineering and Systems, 16*(5), 432–443. https://doi.org/10.22266/ijies2023.1031.38

Syadzily, M. H., Jondri, & Muslim, K. (2024). *Prediksi retweet berdasarkan fitur pengguna, konten, dan waktu menggunakan metode klasifikasi ANN-Cat Swarm Optimization*.

Zannuar, J. M. S. S., & Lhaksmana, K. M. (2021). *Prediksi retweet berdasarkan feature user-based menggunakan metode klasifikasi Random Forest*