KLIK: Kajian Ilmiah Informatika dan Komputer
ISSN 2723-3898 (Media Online)
Vol 4, No 1, Agustus 2023, Hal 124-133
DOI 10.30865/klik.v4i1.1067
https://djournals.com/klik

Retweet Prediction Based on User-Based, Content-Based, and Time-Based Features Using ANN Optimized with GWO

Irgi Aditya Rachman, Jondri, Kemas Muslim L*

[1] Fakultas Informatika, Program Studi Informatika, Telkom University, Bandung, Indonesia
Email: [1]irgi10969@gmail.com, [2]jondri@telkomuniversity.ac.id, [3,*] kemasmuslim@telkomuniversity.ac.id
Email Penulis Korespondensi: kemasmuslim@telkomuniversity.ac.id

Abstract−Social media has emerged as immensely popular and favored platforms among the masses today. Twitter, being one of the most renowned social media platforms, allows users to express themselves through tweet postings. Retweeting is a crucial feature on Twitter, enabling users to disseminate tweets authored by others. In this context, this research aims to predict retweet behavior using User-Based, Content-Based, and Time-Based features, coupled with an Artificial Neural Network classifier optimized with Grey Wolf Optimization. One of the challenges in retweet prediction lies in class imbalance, where the number of retweets on certain tweets is significantly disproportionate compared to others. To address this issue, this study implements undersampling and oversampling techniques. Undersampling reduces the number of samples from the majority class, whereas oversampling involves duplicating or synthesizing samples from the minority class, thereby creating class balance. The research successfully achieves promising results in retweet prediction. After applying oversampling techniques, the classification process attains an accuracy of 85.58%, precision of 87.77%, recall of 83.92%, and F1-score of 85.80%. These results demonstrate the effectiveness of the proposed method in retweet prediction and handling class imbalance issues.

Keywords: Social Media; Twitter; Retweet; Prediction; Imbalance Class; Undersampling; Oversampling; Artificial Neural Network;

1. INTRODUCTION

Social media is a platform where everyone can express themselves. On social media, people can share various things in the form of images and writings. Social media has grown rapidly and become widespread nowadays. Presently, social media is utilized daily by countless individuals, including celebrities, organizations, individuals, and labor unions, among others. These users collectively share millions of messages on diverse subjects such as politics, sports, health, news, technology, and more [1]. Social media has been widely used as one of the most common sources of information [2]. As of today, Twitter is regarded as one of the most popular social media platforms among the general public.

Twitter is one of the most popular social media platforms currently, enabling users to share information through messages known as tweets. Various activities can be performed using Twitter. Users can follow others to receive updates from the accounts they follow, and they can create tweets in the form of text, videos, or images to share with their followers [3].

Twitter has numerous features such as likes, retweets, comments, and others. One interesting feature that emerged on Twitter is the retweet feature, The feature that permits users to disseminate a tweet authored by another user on Twitter is known as "retweeting" [4].Retweets play a crucial role in information dissemination on Twitter. Popular tweets reflect current trends on the platform, and Twitter itself is one of the most significant social media platforms [5]. Exactly, retweet is indeed a unique feature on Twitter. When a user comes across an interesting post uploaded by another user and wants to repost it, they can retweet that post to share it with their own followers. This allows users to amplify and spread content they find engaging or informative, reaching a wider audience beyond the original tweet's reach. It promotes the dissemination of valuable information, ideas, and discussions across the Twitter platform [4]. Structurally, retweeting is similar to email forwarding, where users repost a message originally posted by someone else [6]. When composing a Tweet, it is expected that its content will have a significant impact on users and offer them valuable information. Twitter posts typically contain hashtags, URLs, titles, and other elements, constrained by a character limit of 280. Therefore, it is crucial to concentrate on the content and adhere to the character limit to create a favorable impression and generate increased reader interest. Users can express their satisfaction by using the 'like' and 'retweet' functionalities on the original posts. Predicting the number of 'likes' and 'retweets' a post might receive can assist users in enhancing their content and engaging a larger audience. According to researchers, a Tweet gains popularity when it receives a higher number of retweets and likes from the user community who find the information conveyed in the post relevant and engaging [3]. With such interconnections among users, the author assumes that the number of likes, tweets, and followers has a significant influence and a strong correlation with the number of retweets. The larger the number of followers and retweeters a user has, the higher the likelihood of their tweet being retweeted [7]. This phenomenon results in an information dissemination gap on Twitter, where only a few tweets receive a significant number of retweets compared to other tweets.

Researchers will slightly build upon the previous study related to this research. In 2022, Jondri conducted a study entitled "Retweet Prediction on Twitter Using Decision Tree Method," where the aim was to predict retweets using the Decision Tree classification method [8]. In 2018, Hoang et al. conducted a study to predict whether a Twitter post would be retweeted by other users and to forecast the dissemination of tweets using user-based, content-based, and time-based features [7]. In 2020, Daga et al. conducted a study to predict likes and retweets using several different machine-learning classification methods and employed two distinct text processing approaches utilized in Natural Language Processing (NLP) [3]. In 2021, Surbhi et al. conducted research to study the prediction of retweets as a function of value-based

systems. Their study also included a comparative experimental analysis with the features used in the previous research, namely value-based features. The experimental results using different machine learning algorithms indicated that the value-based system outperformed emotions, sentiments, and specific topic emotions in predicting user retweet decisions [9].

In this research, the investigators aim to predict the likelihood of retweets being retweeted by users using the classification method of Artificial Neural Network (ANN) optimized with the Grey Wolf Optimization (GWO) algorithm. Artificial Neural Network (ANN) is a computational concept inspired by the information processing mechanisms observed in biological neural systems, much like the human brain. The fundamental aspect of this concept is a unique structure of the information processing system, consisting of multiple interconnected processing elements (neurons) that work together to address particular problems. ANN possesses learning capabilities akin to the neural system of the human brain, acquiring knowledge through learning from examples. The development of Artificial Neural Networks (ANNs) involves research in the fields of software simulations based on conventional von Neumann computers, as well as hardware simulations, such as indirect implementations of ANNs based on electronics and photonics, and the direct growth of ANNs with biological neuron cells [10]. The term "Artificial Neural Networks" (ANNs) suggests that their creation aimed to develop synthetic systems by emulating the computational principles employed by the nervous system [11].

The topic to be addressed by the author is about predicting whether a tweet will be retweeted or not, based on selected features: user-based, content-based, and time-based, utilizing the Artificial Neural Network machine learning approach, which will be optimized using the Grey Wolf Algorithm. The scope of this research is limited to data collected and used with the keyword “Valorant”.

The purpose of this research is to develop a retweet prediction system using user-based and content-based features with the ANN-GWO classification method. It is expected that the application of the ANN-GWO method will yield high accuracy in predicting retweets for a tweet.

2. RESEARCH METHODOLOGY

The system to be developed is Retweet Prediction based on User-Based, Content-Based, and Time-Based Features Using ANN Classification Optimized with Greywolf Method. The following is the flowchart of the system design that will be constructed.

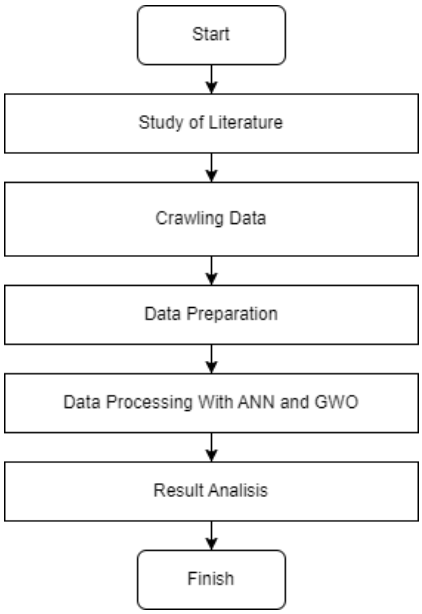

Figure 1. Flowchart System

2.1 Crawling Data

In this phase, the researcher performed data crawling using Netlytic to collect data for this study. The collected data includes User-Based, Content-Based, and Time-Based features. Specifically, approximately 2500 tweets related to the topic "Valorant" were gathered within the timeframe of June 5 to June 12, 2023. The utilization of User-Based, Content-Based, and Time-Based features in data collection allows the researcher to obtain more comprehensive and enriched information. The User-Based feature considers user-related aspects associated with the tweets, such as the number of followers, the number of accounts followed, and user activity. The Content-Based feature involves text analysis of the tweets, including the used keywords, tweet length, and the utilization of specific languages. On the other hand, the Time-Based feature focuses on the publication time of tweets, thereby aiding in studying trends and user behavior patterns over time.

2.2 Data Preparation

After obtaining the data, the next step is data processing to prepare it for the model. One critical stage is checking the class imbalance in the acquired data. Class 0 represents tweets that were not retweeted, while class 1 represents retweeted tweets. In the obtained dataset, there is a class imbalance where class 0 has 854 data points, whereas class 1 has 1646 data points. Figure 4 shows the visualization of the class imbalance check result. To address the class imbalance issue, several testing scenarios can be employed. Firstly, without handling the class imbalance, meaning using the data as it is. In this scenario, the model will be trained using all available data, but this may lead the model to predominantly predict the majority class and disregard the minority class. The second scenario involves using the undersampling technique using the RandomUnderSampler method. In this scenario, the number of samples from the majority class (class 1) will be randomly reduced to balance the number of samples from both classes. By performing undersampling, the model can pay more attention to the minority class and enhance its ability to predict retweets. The third scenario entails utilizing the oversampling technique using the RandomOverSampler method. In this scenario, the number of samples from the minority class (class 0) will be duplicated or synthetically recreated randomly to balance the number of samples from both classes. By performing oversampling, the model can have more examples to learn the characteristics of the minority class and reduce bias towards the majority class. In this research, testing is conducted using all three scenarios to observe a comparison of the model's performance. This will provide a deeper insight into the influence of class imbalance handling on retweet prediction. Furthermore, model performance evaluation will be carried out using precision, recall, F1-score, and accuracy.

2.3 ANN Classification

Artificial Neural Network (ANN) has been an intriguing research focus in the field of artificial intelligence since the 1980s [12]. The Artificial Neural Network (ANN) is a computational model inspired by the data processing methods observed in biological nervous systems, like the human brain. ANN functions as a computational model designed to imitate the structural and functional characteristics of biological neural networks. Different applications of ANN can be classified either as classification models or regression models [13]. Usually, ANNs are comprised of numerous basic processing units interconnected in a sophisticated communication network. Each unit, or node, represents a simplified version of a biological neuron, capable of transmitting signals or "firing" when it receives a sufficiently strong input from other interconnected nodes [10].

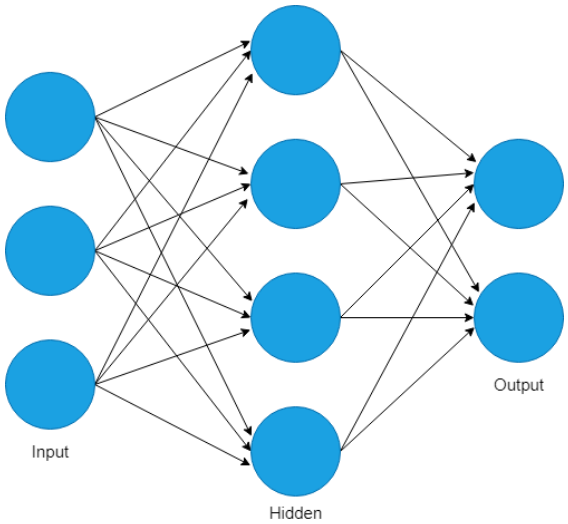

Figure 2. ANN Structure

As seen in Figure 2, ANN consists of 3 layers comprising the input layer, hidden layer, and output layer. ANN is also a popular algorithm in machine learning that aids in classification, clustering, pattern recognition, and prediction in various scientific disciplines [14].

ANN is composed of multiple nodes, interconnected through links, which imitate the behavior of neurons in the human brain. Every link is assigned a weight and can undergo learning processes by adjusting its weight values .To determine how the weights W connect between neurons, the index of the input layer neurons is denoted by i, and the output layer neurons by j. In mathematical notation, considering each input value xi and each hidden value hj, the formula for computing one hj will be as provided in equation 1.

$$h_j = f(b_j + \sum_{i=1}^{n} W_i x_i) \quad (1)$$

In this given context, the symbol f() denotes the activation function (e.g., sigmoid), n represents the count of input features, and bj signifies the bias associated with hidden layer neuron j . Every neuron obtains inputs that are weighted from other neurons and communicates its output to other neurons using the activation function, which is represented by the sigmoid function [15].

2.4 GWO (Grey Wolf Optimizer)

Grey Wolf Optimization (GWO) is an optimization technique based on the hierarchical behavior and social intelligence of wolves [16]. The algorithm attains optimization by mathematically emulating the tracking, encircling, hunting, and attacking behaviors observed in gray wolves. The hunting process of gray wolves consists of three key steps: establishing social hierarchy stratification, encircling the prey, and finally, launching the attack on the prey [17]. The GWO algorithm imitates the hierarchical leadership and hunting behaviors of gray wolves found in nature. It utilizes four types of gray wolves (alpha, beta, delta, and omega) to replicate the leadership hierarchy. Furthermore, the algorithm incorporates three main hunting steps, namely the search for prey, encircling the prey, and attacking the prey. The researchers selected this algorithm based on their findings, which demonstrated that GWO yields remarkably competitive results when compared to widely recognized heuristics like PSO, GSA, DE, EP, and ES [18]. The GWO algorithm draws its primary inspiration from the social hierarchy observed in packs of gray wolves. Within each pack, there exists a system of interactions that establishes dominance and strength. The alpha wolf holds the most significant influence in hunting, eating, and migration, acting as the guiding force for the entire group. In the alpha's absence due to illness or death, the beta wolf, the second strongest, assumes leadership. The omega and delta wolves, with lower influence, hold subordinate roles compared to the alpha and beta. The GWO algorithm seeks to emulate this form of social intelligence exhibited by gray wolves in nature [19].

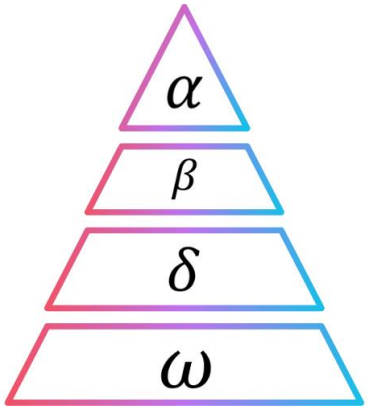

Figure 3. Social Hierarki Koloni Grey Wolf

3. RESULT AND DISCUSSION

3.1 Dataset Description

In this section, the researchers will discuss the dataset used in this study. The dataset was obtained through data crawling using Netlytic. The researchers focused on tweets related to the topic "Valorant" within the timeframe of June 5 to June 12, 2023. This dataset encompasses various features, including User-based features, Content-based features, and Time-based features.

The User-based features encompass information about the users who post tweets, such as the number of followers and the number of accounts they follow. These features can provide insights into the popularity and activity of users, which may influence the likelihood of retweets. The Content-based features involve analyzing the text of the tweets themselves. The researchers analyze keywords used in tweets and the tweet length as important features in predicting retweets. The Time-based features enable the researchers to obtain information about the publication time of tweets. By considering trends and user behavior patterns over time, this feature can provide additional insights in predicting retweets.

The features used are as follows:

a. User Based
   1. user_statuses_count, is the number of tweets from the account.
   2. user_followers_count, is the number of followers.
   3. User_friends_count, is the number of followers.
   4. username_len, is the username length.
   5. age_of_account, is the account age.
   6. Aver_tweets_per_day, is the average tweet that users upload per day.

b. Content Based
   1. has_hashtag, is the post hashtag.
   2. has_url, is a tweet that contains the URL of the content.

3. has_uppercase, are tweets that contain capital letters.
4. has_video, is a tweet that contains a video.
5. has_image, are tweets that contain images.
6. Contain_user_mentioned, are tweets containing "@".
7. has_RT, are tweets containing "RT".
8. opt_length, is the length of the tweet between 70-100 characters.
9. text_length, is the length of the tweet.
10. has_exclamation, are tweets that contain an exclamation point.
11. favorite_count, is the number of post likes.
12. retweet_count, is the number of post retweets.

c. User Based
1. is_posted_in_noon, is a tweet made at 11 – 1 pm.
2. is_posted_on_holiday, is a tweet made on a holiday.
3. is_posted_in_eve, is a tweet made at 4 pm – 9 pm.
4. is_posted_on_weekend, is a tweet made on the weekend.

3.2 Test Scenario

In this research, the researchers applied three testing scenarios to analyze the impact of handling class imbalance on the performance of the retweet prediction model. The three scenarios are as follows:

3.2.1 Scenario 1: Modeling without Class Imbalance Handling

In this scenario, the researchers used the unbalanced dataset without addressing its imbalance. The data was divided into train and test sets with an 80:20 ratio. The objective of this scenario is to evaluate the performance of the retweet prediction model on an unbalanced class dataset.

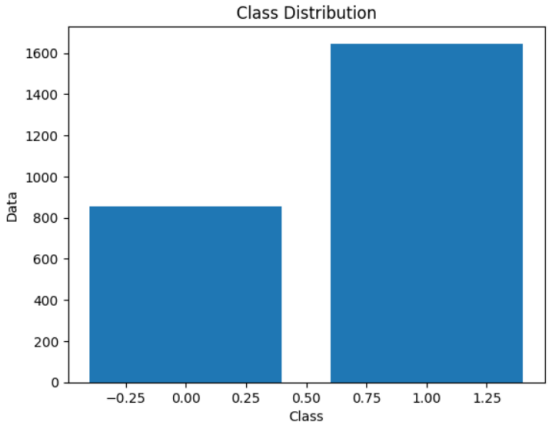

Figure 4. Class Distribution

3.2.2 Scenario 2: Modeling with Oversampling using RandomOverSampler

In this scenario, the researchers applied the oversampling technique using the RandomOverSampler method from the imblearn library. The aim is to increase the number of samples in the minority class to balance it with the majority class. After oversampling, the researchers implemented the ANN-GWO model on the processed dataset.

Figure 5. Class Distribution After Oversampling

3.2.3 Scenario 3: Modeling with Undersampling using RandomUnderSampler

In this scenario, the researchers applied the undersampling technique using the RandomUnderSampler method from the imblearn library. The aim is to reduce the number of samples in the majority class to balance it with the minority class. After undersampling, the researchers implemented the ANN-GWO model on the processed dataset.

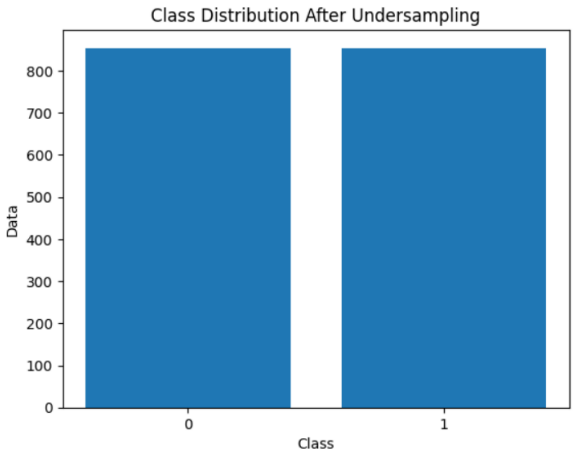

Figure 6. Class Distribution After Undersampling

3.3 Performance Evaluation

Confusion Matrix is a tool used to evaluate the performance of predictive models and conduct measurements in various scenarios. Several metrics are utilized to measure retweet predictions. The assessment is conducted using a confusion matrix, which offers details on the classifications of True Positive, True Negative, False Positive, and False Negative [20]. To evaluate the performance of the model in each testing scenario, the researchers used the following metrics:

a. Accuracy : The proportion of tweets correctly classified by the model.
b. Precision : The proportion of tweets correctly predicted as retweets out of all tweets predicted as retweets.
c. Recall : The proportion of tweets correctly predicted as retweets out of all tweets that are actually retweets.
d. F1-Score : The F1-score is the combined measure of precision and recall that provides an overall view of the model's performance.

Table 1. Confusion Matrix

	Prediction Class	
	True	False
True	TP	TN
False	FP	FN

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

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (3)$$

$$Precison = \frac{TP}{TP + FP} \times 100\% \quad (4)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

The testing was performed five times for each testing scenario using the Artificial Neural Network (ANN) model and the ANN model optimized using the GreyWolf Optimization algorithm (ANN-GWO). Thus, each testing scenario provided accuracy, precision, recall, and F1-Score values representing the model's performance on the processed dataset according to that specific scenario.

3.4 Test Results

The researchers will analyze the testing results in each testing scenario to gain insights into the effectiveness of class imbalance handling techniques and the impact of optimization using the GWO algorithm on model performance.

3.4.1 Scenario 1: Modeling without Class Imbalance Handling

In the first testing using the unbalanced dataset, the data was split into an 80:20 ratio for training and testing purposes. This testing aimed to determine the values of Accuracy, Precision, Recall, and F1-Score on the User-based, Content-

based, and Time-based datasets. The testing was conducted five times using both ANN and ANN optimized by GWO to ensure obtaining maximum coherence in the results.

Table 2. Scenario 1 Test Results

	ANN	*ANN-GWO*
Accuracy	81.20%	82.60%
Precison	82.04%	85.55%
Recall	91.89%	88.89%
F1-Score	86.69%	87.19%

From Table 2, it can be observed that the first testing results obtained from the default ANN model yielded relatively good performance with an accuracy of 81.20%, precision of 82.04%, recall of 91.89%, and F1-score of 86.69%.

Table 3. Hyperparameter Best model Scenario 1

Best Parameters	
hidden_layer_sizes	(11, 11, 11, 11)
activation	*tanh*
solver	*adam*
alpha	7.08-06
learning_rate_init	0.00

Table 3 presents the best parameters obtained for the search of the ANN model optimized with GWO. In scenario 1, the model has four hidden layers, each with 11 neurons. The selected activation function is hyperbolic tangent (tanh). The solver chosen for this scenario is 'adam'. The given value of 'alpha' in this case is 7.08-06, which indicates a very small regularization rate, suggesting that the regularization applied to this model is relatively weak. The value of 'learning_rate_init' given is 0.00, which controls the step size taken at each iteration during model training. By utilizing these best parameters, the F1-Score, accuracy, and precision values have improved from the default values, increasing the F1-Score from 86.69% to 87.19%, accuracy from 81.20% to 82.60%, and precision from 82.04% to 85.55%. However, in scenario 1, there is a decrease in the recall value from 91.89% to 88.89%. Although specific handling for class imbalance has not been applied yet, the model still shows relatively good results in predicting retweets.

3.4.2 Scenario 2: Modeling with Oversampling using RandomOverSampler

In the second testing, the dataset has been oversampled using the RandomOverSampler to balance the class distribution. The data was split into an 80:20 ratio for training and testing purposes. This testing aimed to determine the values of Accuracy, Precision, Recall, and F1-Score on the User-based, Content-based, and Time-based datasets. The testing was conducted five times using both the ANN and the ANN optimized by GWO to ensure obtaining maximum coherence in the results.

Table 4. Scenario 2 Test Results

	ANN	*ANN-GWO*
Accuracy	82.85%	85.58%
Precison	82.44%	87.77%
Recall	85.09%	83.92%
F1-Score	83.74%	85.80%

From Table 4, it can be observed that the first testing results obtained from the default ANN model with oversampled data yielded relatively good performance with an accuracy of 82.85%, precision of 82.44%, recall of 85.09%, and F1-score of 83.74%.

Table 5. Hyperparameter Best model Scenario 2

Best Parameters	
hidden_layer_sizes	(50, 50, 50, 50)
activation	*tanh*
solver	*adam*
alpha	6.27-06
learning_rate_init	0.01

In Table 5, the results of the best parameters obtained for searching the best ANN model optimized with GWO are presented. In scenario 2, using these parameters, the neural network will be trained with the Adam algorithm using the hyperbolic tangent activation function, with four hidden layers, each having 50 neurons. The initial learning rate is set to 0.01, and alpha is 6.27-06. By utilizing these best parameters, the F1-Score, accuracy, and precision values have improved

from the default values, increasing the F1-Score from 83.74% to 85.80%, accuracy from 82.85% to 85.58%, and precision from 82.44% to 87.77%. However, in scenario 2, there is a decrease in the recall value from 85.09% to 83.92%.

3.4.3 Scenario 3: Modeling with Undersampling using RandomUnderSampler

In the third testing, the dataset has been undersampled using the RandomUnderSampler to balance the class distribution. The data was split into an 80:20 ratio for training and testing purposes. This testing aimed to determine the values of Accuracy, Precision, Recall, and F1-Score on the User-based, Content-based, and Time-based datasets. The testing was conducted five times using both the ANN and the ANN optimized by GWO to ensure obtaining maximum coherence in the results.

Tabel 6. Scenario 3 Test Results

	ANN	*ANN-GWO*
Accuracy	78.95%	83.92%
Precison	73.74%	80.00%
Recall	84.08%	86.62%
F1-Score	78.57%	83.18%

From Table 6, it can be observed that the first testing results obtained from the default ANN model with undersampled data yielded relatively good performance with an accuracy of 78.95%, precision of 73.74%, recall of 84.08%, and F1-score of 78.57%.

Table 7. Hyperparameter Best model Scenario 3

Best Parameters	
hidden_layer_sizes	(10, 10, 10, 10)
activation	*tanh*
solver	*adam*
alpha	0.0
learning_rate_init	0.00

In Table 7, the results of the best parameters obtained for searching the best ANN model optimized with GWO are presented. In scenario 3, using these parameters, the neural network will be trained with the Adam algorithm using the hyperbolic tangent activation function, with four hidden layers, each having 10 neurons. The initial learning rate is set to 0.00, and alpha is set to 0.0, indicating that the model learns without any restrictions on its learning rate. By utilizing these best parameters, all the performance metrics improved compared to the default values. The accuracy increased from 78.95% to 83.92%, precision from 73.74% to 80.00%, recall from 84.08% to 86.62%, and F1-score from 78.57% to 83.18%.

3.5 Discussion

Based on the testing results from the three scenarios, the following discussions can be made:

a. Scenario 1 : Scenario 1 shows that the ANN-GWO model produces slightly better performance than the ANN model on the unbalanced dataset. This indicates that optimization using the GWO algorithm can provide improvements in predicting retweets.
b. Scenario 2 and 3 : Scenarios 2 and 3 demonstrate that oversampling and undersampling techniques can improve the model's performance in predicting retweets. However, it is important to note that oversampling tends to yield better results compared to undersampling in terms of accuracy, precision, recall, and F1-Score.

The optimization using the GWO algorithm in the ANN-GWO model consistently improves performance in each testing scenario. This indicates that the GWO algorithm is effective in optimizing the ANN model for predicting retweets. Based on the results of the three different scenarios, the scenario with the dataset oversampled using RandomOverSampler and optimized with the GWO algorithm showed the best performance. The ANN-GWO model in this scenario achieved an accuracy of 85.58%, precision of 87.77%, recall of 83.92%, and F1-score of 85.80%. This indicates that with the use of oversampling and optimization using GWO, the ANN model can achieve better performance in classifying data with imbalanced class distribution. The other scenarios, namely the scenario with the unbalanced dataset and the scenario with the undersampled dataset, also showed performance improvements after being optimized with GWO, but their performance was lower compared to the oversampled dataset scenario. These findings can serve as important guidance in developing more accurate and reliable retweet prediction systems on social media platforms like Twitter. Furthermore, these findings provide insights into the importance of handling class imbalance in retweet prediction modeling. Oversampling and undersampling techniques can be used as effective approaches to address this issue. However, further evaluation and research are needed to deepen the understanding of the impact of class imbalance handling techniques on model performance in a broader context.

This research provides a significant contribution to understanding and enhancing retweet prediction capabilities in the context of social media, particularly on the Twitter platform. By utilizing oversampling and undersampling techniques, the researchers successfully balanced the class distribution in the dataset and achieved more accurate prediction results.

The findings of the study can serve as a foundation for developing more effective retweet prediction algorithms that can be applied in various fields, including digital marketing, social media analysis, and user behavior understanding.

Furthermore, this research highlights the importance of User-based, Content-based, and Time-based features in predicting retweets. These features provide valuable insights into the characteristics of tweets and users that can influence the level of interaction and retweets from other users. The results of the study indicate that considering the combination of these features can significantly enhance the predictive capabilities of the model.

4. CONCLUSION

Social media is a platform where individuals can express themselves. Twitter is one of the most popular social media platforms today, enabling users to communicate and express themselves through tweets, and retweeting allows content to be disseminated more widely by other users. This research aims to predict retweets using various features, such as User-Based, Content-Based, and Time-Based, and optimize them using the Artificial Neural Network (ANN) classification method and GreyWolf algorithm. One of the challenges in predicting retweets is class imbalance, where some tweets have a large number of retweets, while most others have only a few retweets. To address this issue, the study implements undersampling and oversampling techniques. Undersampling reduces the number of samples from the majority class, while oversampling involves duplicating or synthesizing samples from the minority class, creating better class balance. The results of this research show that the proposed approach significantly succeeds in predicting retweets. By implementing oversampling, the classification accuracy increases to 85.58%, precision reaches 87.77%, recall reaches 83.92%, and F1-score reaches 85.80%. This means that the generated model effectively predicts which tweets have the potential to receive more retweets. However, it should be noted that oversampling techniques can also pose the risk of overfitting the model. Therefore, this research should also be balanced with careful testing and further model validation to ensure good generalization in various situations and different datasets. Moreover, continuous efforts to improve and develop the model and adapt to dynamic changes in social media and user behavior will be crucial to ensuring its sustainability.

REFERENCES

[1] D. Henry, E. Stattner, and M. Collard, "Social media, diffusion under influence of parameters : survey and perspectives," *Procedia Comput. Sci.*, vol. 109, pp. 376–383, 2017, doi: 10.1016/j.procs.2017.05.404.

[2] M. Broersma and T. Graham, "TWITTER AS A NEWS SOURCE: How Dutch and British newspapers used tweets in their news coverage, 2007–2011," *Journal. Pract.*, vol. 7, no. 4, pp. 446–464, Aug. 2013, doi: 10.1080/17512786.2013.802481.

[3] I. Daga, A. Gupta, R. Vardhan, and P. Mukherjee, "Prediction of Likes and Retweets Using Text Information Retrieval," *Procedia Comput. Sci.*, vol. 168, pp. 123–128, 2020, doi: 10.1016/j.procs.2020.02.273.

[4] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network," in *2010 IEEE Second International Conference on Social Computing*, Minneapolis, MN, USA: IEEE, Aug. 2010, pp. 177–184. doi: 10.1109/SocialCom.2010.33.

[5] A. Kupavskii *et al.*, "Prediction of retweet cascade size over time," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, Maui Hawaii USA: ACM, Oct. 2012, pp. 2335–2338. doi: 10.1145/2396761.2398634.

[6] D. Boyd, S. Golder, and G. Lotan, "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter," in *2010 43rd Hawaii International Conference on System Sciences*, Honolulu, HI: IEEE, Jan. 2010, pp. 1–10. doi: 10.1109/HICSS.2010.412.

[7] T. B. N. Hoang and J. Mothe, "Predicting information diffusion on Twitter – Analysis of predictive features," *J. Comput. Sci.*, vol. 28, pp. 257–264, Sep. 2018, doi: 10.1016/j.jocs.2017.10.010.

[8] J. Jondri, "PREDIKSI RETWEET PADA TWITTER MENGGUNAKAN METODE DECISION TREE," *CSRID Comput. Sci. Res. Its Dev. J.*, vol. 14, no. 2, p. 113, Sep. 2022, doi: 10.22303/csrid.14.2.2022.113-124.

[9] S. Kakar, D. Dhaka, and M. Mehrotra, "Value-Based Retweet Prediction on Twitter," *Informatica*, vol. 45, no. 2, Jun. 2021, doi: 10.31449/inf.v45i2.3465.

[10] Q. Zhang, H. Yu, M. Barbiero, B. Wang, and M. Gu, "Artificial neural networks enabled by nanophotonics," *Light Sci. Appl.*, vol. 8, no. 1, p. 42, May 2019, doi: 10.1038/s41377-019-0151-0.

[11] A. M. Zador, "A critique of pure learning and what artificial neural networks can learn from animal brains," *Nat. Commun.*, vol. 10, no. 1, p. 3770, Aug. 2019, doi: 10.1038/s41467-019-11786-6.

[12] Y. Chen, L. Song, Y. Liu, L. Yang, and D. Li, "A Review of the Artificial Neural Network Models for Water Quality Prediction," *Appl. Sci.*, vol. 10, no. 17, p. 5776, Aug. 2020, doi: 10.3390/app10175776.

[13] P. G. Asteris and V. G. Mokos, "Concrete compressive strength using artificial neural networks," *Neural Comput. Appl.*, vol. 32, no. 15, pp. 11807–11826, Aug. 2020, doi: 10.1007/s00521-019-04663-2.

[14] B. Eftekhar, K. Mohammad, H. E. Ardebili, M. Ghodsi, and E. Ketabchi, "Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data," *BMC Med. Inform. Decis. Mak.*, vol. 5, no. 1, p. 3, Dec. 2005, doi: 10.1186/1472-6947-5-3.

[15] C. Surianarayanan, J. J. Lawrence, P. R. Chelliah, E. Prakash, and C. Hewage, "Convergence of Artificial Intelligence and Neuroscience towards the Diagnosis of Neurological Disorders—A Scoping Review," *Sensors*, vol. 23, no. 6, p. 3062, Mar. 2023, doi: 10.3390/s23063062.

[16] Department of Mathematics, Graphic Era Deemed to be University, Dehradun, India *et al.*, "Optimization of Complex System Reliability using Hybrid Grey Wolf Optimizer," *Decis. Mak. Appl. Manag. Eng.*, vol. 4, no. 2, pp. 241–256, Oct. 2021, doi: 10.31181/dmame210402241n.

[17] M. H. Nadimi-Shahraki, S. Taghian, and S. Mirjalili, "An improved grey wolf optimizer for solving engineering problems," *Expert Syst. Appl.*, vol. 166, p. 113917, Mar. 2021, doi: 10.1016/j.eswa.2020.113917.

[18] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey Wolf Optimizer," *Adv. Eng. Softw.*, vol. 69, pp. 46–61, Mar. 2014, doi: 10.1016/j.advengsoft.2013.12.007.
[19] Q. Al-Tashi, H. Md Rais, S. J. Abdulkadir, S. Mirjalili, and H. Alhussian, "A Review of Grey Wolf Optimizer-Based Feature Selection Methods for Classification," in *Evolutionary Machine Learning Techniques*, S. Mirjalili, H. Faris, and I. Aljarah, Eds., in Algorithms for Intelligent Systems. Singapore: Springer Singapore, 2020, pp. 273–286. doi: 10.1007/978-981-32-9990-0_13.
[20] E. T. Arifin, J. Jondri, and I. Indwiarti, "Prediction Retweet Using User-Based and Content-Based with ANN-GA Classification Method," *Build. Inform. Technol. Sci. BITS*, vol. 4, no. 2, pp. 522–528, Sep. 2022, doi: 10.47065/bits.v4i2.1931.