1. Introduction

The development of technology directs humans to a digital-based life. The presence of digital technology has a major impact on human life in various aspects. One of them is the aspect of giving and getting information. In this digital era, social networks are the main intermediary in the diffusion of information [1]. In social networks, humans or users are connected through social relationships, personal attachment, neighborhood, and other factors [2]. Many people spend hours online getting information about products [3], news, hobbies, and interests [4]. This has led to increased interactions on social networks. In addition, people on social media can also share information independently. Thus, the information can spread quickly over a wide area [5].

Microblogs such as Twitter are one of the media in the digital era that can disseminate information widely and quickly. On Twitter, the main mechanism for the diffusion of information is retweeting [6]. Retweeting is an activity to repost a tweet that looks like the original tweet. Not only limited to reposting but retweeting also includes quote retweets, which is retweeting with comments [7]. The more retweets of a tweet, the more widespread the information becomes. So, studying the spread of information through retweet prediction can assist in determining whether or not a tweet has the potential to go viral.

Our research on retweet prediction is based on several studies related as a reference. Research conducted by Bunyamin and Tunys by comparing various machine learning methods such as Passive Aggressive, Linear Support Vector Machine, Logistic Regression, Decision Tree, and Random Forest. The features used in this research are user and tweet-based. The results of the study indicate that the Random Forest model achieves the highest performance of the other learning methods considered. Then the use of user-based features and tweets outperforms user only and tweets only features, which means that these two features affect retweet predictions [8].

Other research on retweet prediction has also been carried out by Hoang and Mothe. This study uses a Random Forest algorithm and features based on user, content, and time. This study produces an evaluation score between 70% to 82% according to the data using the F-measure metric. This study also shows that some features have a greater influence on predicting retweets such as the number of followers, number of followees, and number of group users belong. In addition, time-based features are also highly correlated with retweet ability [9].

There are similar studies on retweet prediction that consider several implicit features, such as the research of Hoang and Mothe and the research of Daga, et al. Hoang and Mothe use the sentiment feature as an implicit feature, while Daga et al. implement information retrieval to predict retweets. Daga et al. compared two text vectorization methods, namely TF-IDF weighting and Doc2Vec with various learning algorithms such as Random Forest, Support Vector Machine, Neural Network, Logistic Regression, and Multinomial Naïve Bayes. The results show that all models provide 10% to 15% better accuracy using the TF-IDF method than Doc2Vec [4].

In this study, we aim to build a model that can predict which tweet will be retweeted and not retweeted. The features used are a combination of user-based, content-based, and time-based features from previous studies. In addition, we also consider the text as a feature by using TF-IDF weighting. The novelty of this study is we implementing a binary classification method, namely Evolutionary Undersampling Boosting. We consider our scenario to be a data imbalance problem because the number of instances in the class of not being retweeted greatly outnumbers the number of instances in the class of being retweeted.