Abstract

Twitter is a social media where users can post information in the form of tweets. Tweets posted by users can be re-shared by other users via retweets, tweets that have multiple retweets spread more quickly and widely than those that don't get retweets. But not all tweets get retweets because there are factors that influence them. In this study, the features used are user-based, time-based and content-based, as well as using a classification and regression tree (CART) decision tree machine learning algorithm to predict the retweet class. The problem faced in this research is unbalanced data, then overcome it by oversampling and undersampling. After overcoming unbalanced data, the performance of the model increased, especially when undersampling for max_depth = 4 resulted in 85% accuracy and f1.

Keywords: retweet, tweet, DT, CART, oversampling, undersampling