Abstract

This research analyzes the sentiment reviews of the SIREKAP application on the Google Play Store using two machine learning algorithms, namely Naïve Bayes and Support Vector Machine (SVM). The dataset used consists of 19,925 reviews that have gone through preprocessing stages, including text cleaning, stopword removal, stemming, and tokenization. To overcome data imbalance, oversampling and undersampling techniques were applied. Furthermore, TF-IDF is used for feature extraction, converting text into numerical representation. The dataset is divided into 80% training data (15,940 data) and 20% test data (3,985 data). The results show that oversampling provides better performance than undersampling. In the oversampling method, the SVM algorithm achieved the highest accuracy of 95%, with consistent precision, recall, and F1-score values across all sentiment classes. The Naïve Bayes algorithm also performed quite well, with an accuracy of 77% on the oversampled data. In contrast, in the undersampling method, both algorithms have the same accuracy of 61%. This study confirms that the combination of oversampling technique and SVM algorithm is the best approach to handle imbalanced data and provides important insights into user perception of the SIREKAP application.