ABSTRACT

Previous research on sentiment classification has been limited to small-scale datasets and has lacked a comprehensive comparison of various oversampling methods. Data imbalance poses a significant challenge in sentiment analysis, where the dominance of the majority class causes classification models to be biased and overlook the minority class. This study aims to enhance classification algorithms by applying three oversampling techniques: Synthetic Minority Oversampling Technique (SMOTE), Borderline-SMOTE, and Random Oversampling. The dataset used consists of 10,000 Tokopedia user reviews collected through web scraping. Three classification algorithms—Random Forest, Naïve Bayes, and Support Vector Machine (SVM)—are employed for model training. Model performance is evaluated using Accuracy, Precision, Recall, F1-Score, and Area Under the Curve (AUC) metrics. This research is the first comparative study to apply three oversampling methods on a large-scale dataset (10,000 reviews) using three different algorithms in the context of Indonesian e-commerce. The results show that the combination of SMOTE and the Naïve Bayes algorithm yields stable performance, achieving an Accuracy of 86.71%, Precision of 80.43%, Recall of 70.38%, F1-Score of 75.07%, and AUC of 90.62%. These findings demonstrate that SMOTE is effective in improving classification performance.

Keywords: Sentiment Analysis, Imbalanced Data, Oversampling, Random Forest, Naïve Bayes, Support Vector Machine.