Abstract

Sentiment Analysis is a field of Natural Language Processing (NLP) that aims to classify text into positive, negative, or neutral sentiments. Several obstacles prevent the model from working optimally, such as imbalanced dataset, when a dataset has several classes with fewer samples. To overcome this, several methods involve oversampling and data augmentation to add more data to the minority class. Additionally, cost-sensitive learning is where the model handles the weight of each data class without adding more data with the change from Cross-Entropy loss function to Focal Loss and Adam optimizer to NAdam. Another obstacle is that with the latest BERT-based models discovered are large models. These models take a longer time to process and therefore require enormous computing power. In this study, these methods evaluated on an imbalanced Indonesian General Sentiment Analysis dataset that consists of informal sentences to measure the performance of BERT, IndoBERT, DistilBERT, DistilBERT-Indo, and RoBERTa in handling imbalanced datasets and the time taken. As a result, random oversampling outperforms data augmentation, and IndoBERT outperforms other models compared with 88.46% Macro F1-Score and 94.41% on the boosted dataset, with Focal Loss and NAdam were on average better than their respective counterparts of Cross-Entropy and Adam.

Keywords: sentiment analysis, transformers, imbalanced dataset, oversampling, cost-sensitive learning