ABSTRACT

The use of abusive language and hate speech on Indonesian Twitter has recently increased. The existence of these sentences on social media has many negative impacts, causing misunderstandings among netizens, and children can also be exposed to language that is not appropriate for their age, which can cause harm to their psychology. Additionally, abusive language and hate speech often overlap, making the distinction unclear. So it is important to carry out a multi-label classification of abusive language and hate speech, especially on Twitter. In previous research, many have used classical machine learning models for multilabel classification. However, this research has not achieved good accuracy due to the limited ability of the model to represent features, understand context, overcome imbalanced data, and cannot use a transfer learning approach. Based on the author's literature study, there has been no research that uses contextual embedding models such as BERT and RoBERTa with pre-trained models that have been trained in informal Indonesian. Therefore, this research propose the use of new pre-trained models, namely IndoBERTweet, Indonesian RoBERTa Base, IndoRoBERTa small to overcome these shortcomings. The use of different data balancing methods, as well as adjusting hyperparameter tuning, is a challenge to get the best classification accuracy. Based on test results, IndoBERTweet with random oversampling and optimal hyperparameters (learning rate 1e-4, batch size 64, 3 epoch), outperforms other models, namely with an accuracy of 0.86, average precision 0.85, average recall 0.86, and average F1 score 0.85. Additionally, data balancing can be useful in improving accuracy in some scenarios such as random oversampling in the IndoBERTweet model, but its effectiveness is not consistent across models and configurations.

Keywords: Abusive Language, Hate Speech IndoBERTweet, Indonesian RoBERTa Base, IndoRoBERTa Small