ABSTRACT

Online toxicity detection in Indonesian digital interactions poses a significant challenge due to the complexity and nuances of language. This study aims to evaluate the effectiveness of the BERT and RoBERTa language models, specifically IndoBERTweet, IndoBERT, and Indonesian RoBERTa, for identifying toxic content in Bahasa Indonesia. Our research methodology includes data collection, dataset pre-processing, data annotation, and model fine-tuning for multi-label classification tasks. The model performance is assessed using macro average of precision, recall, and F1-score. Our findings show that IndoBERTweet, fine-tuned under optimal hyperparameters (5e-5 learning rate, a batch size of 32, and three epochs), outperforms the other models with a precision of 0.85, recall of 0.94, and an F1-score of 0.89. These findings indicate that IndoBERTweet performs better in detecting and classifying online toxicity in Bahasa Indonesia. The study's implications extend to fostering a safer and healthier online environment for Indonesian users, while also providing a foundation for future research exploring additional models, hyperparameter optimizations, and techniques for enhancing toxicity detection and classification in the Indonesian language.

Keywords: *Online toxicity, Bahasa Indonesia, Multi-label classification, IndoBERTweet, IndoBERT, Indonesian RoBERTa, Macro Average F1-score*