ABSTRACT

With the rise of social media usage in Indonesia, harmful behavior such as cyberbullying has become more common and often goes unnoticed, especially when conveyed in subtle or implicit ways. Detecting such online abuse poses significant challenges, particularly in informal and context-dependent Indonesian-language content. This study explores the use of the RoBERTa transformer model to classify Indonesian social media comments into three categories: explicit cyberbullying, implicit cyberbullying, and non-cyberbullying. A dataset of 15,000 posts was collected and manually labeled, with equal distribution among the three classes. The text underwent a series of preprocessing steps, including normalization and stemming, to ensure consistency and clarity. Multiple experiments were conducted using different combinations of learning rates, batch sizes and optimizers. Although RoBERTa was the main model, the BERT model was also included for performance comparison. Results showed that while both models performed well, BERT achieved the best F1-score of 0.7851 using the Adam optimizer (learning rate 5e-5, batch size 32), slightly outperforming RoBERTa's top F1-score of 0.7764 under the AdamW optimizer (learning rate 1e-5, batch size 32). However, both models faced challenges in handling implicit or sarcastic expressions due to the informal and nuanced nature of the language. These findings highlight the importance of contextual understanding in cyberbullying detection and show that transformer-based models—especially when fine-tuned properly—can be powerful tools for improving online safety.

Keywords: Cyberbullying Detection, Explicit, Implicit, RoBERTa, Indonesian Social Media, NLP, Text Classification