ABSTRACT

The advancement of technology and social media has significantly transformed interaction patterns in Indonesian society, with over 49.9% of active internet users accessing social media daily. This high usage has given rise to the phenomenon of cyberbullying, which proliferates easily due to the vastness of the digital space and minimal regulation. Cyberbullying can cause severe psychological impacts and deteriorate the mental health of victims. Cyberbullying comments often do not appear explicitly but utilize sarcasm, puns, or subtle satire, making them difficult to detect using simple keyword-based matching approaches. Therefore, early detection of cyberbullying comments is a crucial strategy for creating a safe digital environment. This study proposes an Indonesian cyberbullying detection model using the DistilBERT architecture, selected for its computational efficiency compared to standard BERT while maintaining comparable performance. The implementation involves data preprocessing, tokenization, and fine-tuning the DistilBERT model using the AdamW optimizer. Furthermore, an exploration of 24 hyperparameter combinations was conducted varying learning rate, batch size, dropout, and weight decay to obtain the optimal configuration. The best model achieved a validation accuracy of 91.86% with a configuration of learning rate 0.00001, batch size 16, dropout 0.3, and weight decay 0.01. Testing results showed that the best model yielded an accuracy of 88.44%, precision of 88.43%, recall of 88.53%, and F1-score of 88.43%. These results indicate that the DistilBERT model possesses good generalization capabilities in classifying Indonesian cyberbullying comments, thereby potentially supporting the creation of a safer digital ecosystem.

Keywords :cyberbullying detection, DistilBERT, natural language processing, text classification