Abstract

Twitter is one of the social media platforms used to channel user's opinions and it is common for users to misuse this feature to express hate speech. Hate speech, intentional or unintentional, can trigger conflicts, especially on social media platforms with a large user base like Twitter, where hate speech can easily spread and reach its targets. Therefore, a system is needed to detect hate speech to prevent its spread and handle it. In this study, hyperparameter fine-tuning is performed on a pretrained BERT Model classify data containing hate speech with the collected data from Twitter. Adam Optimizer and AdamW Optimizer are used on the constructed BERT model, and their accuracies will be compared to determine which one yields the best result. This research found that the fine tuning the right parameter for each optimizer can result in a significantly higher accuracy compared to the model with default parameter setting. For the task of identifying hate speech in data, BERT model optimized by AdamW optimizer can reach an accuracy of 90.08% by setting the initial learning rate to 1e-5 and initial weight decay to 1e-3, an increase of 40.03% from the baseline. While BERT model optimized with Adam reaches 90.03% in accuracy by setting the initial learning rate to 1e-5 and using its default initial weight decay, an increase of 40.38% from the baseline.

Keywords : Twitter, BERT, Optimizer, Adam, AdamW, Hate Speech