Detecting Hate Speech on Indonesian Twitter Using the IndoBERTweet Method

Bijak Algifan Putra[1], Yuliant Sibaroni[2], Sri Suryani Prasetyowati[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]balgifanp@student.telkomuniversity.ac.id, [2]yuliant@telkimuniversity.ac.id,
[3]srisuryani@telkomuniversity.ac.id

Abstract

Social media is used by the public as a medium for online communication. It is not uncommon for this platform to be used for bad things, namely hate speech. Hate speech on social media must be detected to avoid conflict between citizens and avoid bad examples for social media users. This encourages researchers to create an automatic hate speech detection system, in previous studies conducted using machine learning approaches. But in recent years pre-trained language models have shown a major breakthrough in the field of NLP (Natural language processing). Some studies that have used pre-trained model methods such as detection of hate speech and abusive language on Twitter using Bidirectional Encoder Representations From Transformers (BERT) and detection of the use of abusive sentences in Indonesian text using Indonesia Bidirectional Encoder Representations from Transformers (IndoBERT). In this research, the IndoBERTweet method is used to detect hate speech on Twitter. The creation of the testing and training system involves building a dataset by crawling data and labeling data manually. The IndoBERTweet model has shown excellent performance in detecting hate speech with optimal results. Evaluation results show that through hyperparameter tuning to get the best model, the model achieved an accuracy of 87.1%.

Keywords: hate speech, Indobertweet, pre-trained, hyperparameter