ABSTRAK

Maraknya penggunaan media sosial di Indonesia telah meningkatkan penyebaran konten negatif seperti ujaran kebencian (*hate speech*) dan bahasa kasar (*abusive language*), yang berpotensi memicu konflik sosial. Penelitian ini berfokus pada klasifikasi otomatis kedua jenis konten tersebut menggunakan algoritma *Bidirectional Gated Recurrent Unit* (Bi-GRU), yang pada penelitian sebelumnya menunjukkan potensi untuk dioptimalkan. Dengan *membandingkan tiga model pre-trained word embedding*; *FastText*, IndoBERT, dan IndoBERTweet. Penelitian ini bertujuan menemukan hasil kombinasi paling efektif. Hasil pengujian dengan skema *stratified cross-validation* menunjukkan bahwa IndoBERTweet menjadi model yang paling unggul dengan rata-rata akurasi 75,11%, diikuti oleh IndoBERT dengan 74,88%, dan FastText dengan 74,23% yang membuktikan model yang spesifik pada domain media sosial lebih efektif. Meskipun akurasi tersebut lebih rendah dibandingkan penelitian sebelumnya pada tugas yang lebih sederhana, hasil ini dinilai optimal mengingat kompleksitas klasifikasi multikelas pada dataset yang tidak seimbang tanpa menggunakan metode *resampling*.

Kata kunci: Klasifikasi Teks, *Hate Speech*, *Abusive*, *Bidirectional Gated Recurrent Unit* (Bi-GRU), *Natural Language Processing* (NLP), *FastText*, IndoBERT, IndoBERTweet.

ABSTRACT

The widespread use of social media in Indonesia has increased the spread of negative content such as hate speech and abusive language, which has the potential to trigger social conflict. This research focuses on the automatic classification of these two types of content using the Bidirectional Gated Recurrent Unit (Bi-GRU) algorithm, which previous research has shown potential to optimize. By comparing three pre-trained word embedding models; FastText, IndoBERT, and IndoBERTweet. This research aims to find the most effective combination result. The test results with stratified cross-validation scheme show that IndoBERTweet is the most superior model with an average accuracy of 75.11%, followed by IndoBERT with 74.88%, and FastText with 74.23% which proves that models specific to the social media domain are more effective. Although the accuracy is lower than previous studies on simpler tasks, this result is considered optimal considering the complexity of multiclass classification on unbalanced datasets without using resampling methods.

Keywords: Text Classifiation, Hate Speech, Abusive, Bidirectional Gated Recurrent Unit (Bi-GRU), Natural Language Processing (NLP), FastText, IndoBERT, IndoBERTweet.