ABSTRAK

Abstrak penelitian ini bertujuan untuk mengembangkan model deteksi ujaran kebencian (*hate speech*) dan bahasa abusive menggunakan algoritma Bidirectional Long Short-Term Memory (BiLSTM). Model ini memanfaatkan embedding IndoBERT, IndoBERTweet, dan FastText untuk menganalisis dataset tweet berbahasa Indonesia dari platform GitHub yang terdiri dari 13.170 tweet yang telah dilabeli. Penelitian ini mengevaluasi kinerja model BiLSTM dalam mengklasifikasikan tweet menjadi empat kategori: (1) *Hate Speech* yang bersifat abusive, (2) Non-*Hate Speech* tetapi bersifat abusive, (3) *Hate Speech* tetapi tidak bersifat abusive, dan (4) Non-*Hate Speech* dan Non-Abusive. Hasil penelitian menunjukkan bahwa model BiLSTM dengan embedding IndoBERTweet memberikan akurasi terbaik, dengan F1-score rata-rata mencapai 73,87%. Model ini menunjukkan potensi besar dalam mendeteksi dan mengklasifikasikan ujaran kebencian dan bahasa abusive secara efektif di media sosial berbahasa Indonesia.

Kata kunci: Deteksi ujaran kebencian, BiLSTM, IndoBERT, IndoBERTweet, FastText.

ABSTRACT

The abstract of this research aims to develop a model for detecting hate speech and abusive language using the Bidirectional Long Short-Term Memory (BiLSTM) algorithm. This model utilizes IndoBERT, IndoBERTweet, and FastText embeddings to analyze a dataset of Indonesian-language tweets from the GitHub platform, consisting of 13,170 labeled tweets. This study evaluates the performance of the BiLSTM model in classifying tweets into four categories: (1) abusive hate speech, (2) non-hate speech but abusive, (3) hate speech but not abusive, and (4) non-hate speech and non-abusive. The results show that the BiLSTM model with IndoBERTweet embedding provides the best accuracy, with an average F1-score of 73.87%. This model shows great potential in effectively detecting and classifying hate speech and abusive language on Indonesian-language social media.

Keywords: Hate speech detection, BiLSTM, IndoBERT, IndoBERTweet, FastText.