

ABSTRAK

Perkembangan pesat media sosial meningkatkan interaksi digital yang juga diwarnai komentar *Toxic* yang merusak kualitas komunikasi dan kenyamanan pengguna. Oleh karena itu, deteksi otomatis komentar *Toxic* menjadi kebutuhan penting untuk mendukung pengelolaan konten secara efektif. Penelitian ini bertujuan mengembangkan model klasifikasi komentar *Toxic* serta mengukur akurasi dan performa Long Short-Term Memory (LSTM) dalam klasifikasi multi-label menggunakan 10.146 data hasil crawling dan scraping dari Twitter dan Instagram. Metode yang digunakan adalah LSTM dengan pendekatan multi-label. Tahapan penelitian meliputi *preprocessing* data, embedding kata, pelatihan model, serta evaluasi menggunakan metrik akurasi, precision, recall, dan F1-score. LSTM dipilih karena mampu mengingat konteks urutan kata dan menangani data sekuensial panjang secara efektif. Hasil penelitian menunjukkan model mampu memberikan prediksi yang tepat dan andal. Evaluasi menunjukkan metode tanpa *Undersampling* memperoleh micro average F1-score 0,53, macro average 0,41, dan weighted average 0,54, lebih tinggi dibandingkan metode dengan *Undersampling*. Penelitian ini juga diimplementasikan dalam antarmuka web interaktif menggunakan Streamlit yang dapat diakses publik melalui Ngrok.

Kata kunci : Klasifikasi komentar *Toxic*, Klasifikasi multilabel, *Long Short Term Memory*, *Undersampling*.

ABSTRACT

The rapid growth of social media has increased digital interactions, which are often accompanied by Toxic comments that can harm communication quality and user comfort. Therefore, automatic detection of Toxic comments has become essential to support effective content moderation. This study aims to develop a Toxic comment classification model and evaluate the accuracy and performance of the Long Short-Term Memory (LSTM) method in multi-label classification using 10,146 datasets collected through crawling and scraping from Twitter and Instagram. The proposed method applies LSTM with a multi-label approach. The research process includes data preprocessing, word embedding, model training, and evaluation using accuracy, precision, recall, and F1-score metrics. LSTM was chosen for its ability to capture word sequence context and effectively handle long sequential data. The results show that the model can provide accurate and reliable predictions. Evaluation indicates that the method without Undersampling achieved a micro-average F1-score of 0.53, a macro-average of 0.41, and a weighted-average of 0.54, outperforming the Undersampling approach. Furthermore, the study was implemented into an interactive web interface using Streamlit, publicly accessible through Ngrok.

Keywords: Toxic comment classification, Multilabel classification, Long Short-Term Memory, Undersampling.