ABSTRAK

Deteksi *online toxicity* dalam interaksi digital di Indonesia menimbulkan tantangan yang signifikan karena kompleksitas dan nuansa bahasa. Studi ini bertujuan untuk mengevaluasi efektivitas model bahasa BERT dan RoBERTa, khususnya IndoBERTweet, IndoBERT, dan Indonesian RoBERTa, untuk mengidentifikasi konten beracun dalam Bahasa Indonesia. Metodologi penelitian kami mencakup pengumpulan data, pra-pemrosesan dataset, anotasi data, dan *fine-tuning* model untuk tugas klasifikasi multi-label. Kinerja model dinilai menggunakan *precision*, *recall*, dan *f1-score*. Temuan kami menunjukkan bahwa IndoBERTweet, yang telah dilakukan *fine-tuning* dengan hyperparameter optimal (*learning rate* sebesar 5e-5, *batch size* sebesar 32, dan tiga *epoch*), unggul dari model lainnya dengan rata-rata makro *precision* sebesar 0.85, *recall* sebesar 0.94, dan *f1-score* sebesar 0.89. Temuan ini menunjukkan bahwa IndoBERTweet lebih baik dalam mendeteksi dan mengklasifikasi *online toxicity* dalam Bahasa Indonesia. Implikasi penelitian ini meluas untuk mendorong lingkungan *online* yang lebih aman dan sehat bagi pengguna Indonesia, juga memberikan dasar untuk penelitian masa depan yang mengeksplorasi model tambahan, optimasi hyperparameter, dan teknik untuk meningkatkan deteksi dan klasifikasi *toxicity* dalam bahasa Indonesia.

Kata Kunci: *Online toxicity,* Bahasa Indonesia, *Multi-label classification, IndoBERTweet, IndoBERT, Indonesian RoBERTa, Macro Average F1-score*