ABSTRAK

Perkembangan teknologi dan media sosial telah mengubah cara masyarakat Indonesia berinteraksi, dengan lebih dari 49,9% pengguna internet aktif menggunakan media sosial setiap hari. Tingginya penggunaan ini memunculkan fenomena *cyberbullying*, yang mudah terjadi karena luasnya ruang digital dan minimnya regulasi. *Cyberbullying* dapat menimbulkan dampak negatif secara psikologis dan memperburuk kesehatan mental korban. Bentuk komentar *cyberbullying* sering kali tidak muncul secara eksplisit, melainkan menggunakan bahasa sarkastik, plesetan, atau sindiran halus, sehingga sulit dideteksi menggunakan pendekatan berbasis pencocokan kata kunci sederhana. Oleh karena itu, deteksi dini komentar *cyberbullying* menjadi strategi penting untuk menciptakan ruang digital yang aman. Penelitian ini mengusulkan model deteksi komentar *cyberbullying* berbahasa Indonesia menggunakan arsitektur DistilBERT, yang lebih efisien secara komputasi dibandingkan BERT namun tetap mempertahankan performanya. Implementasi model DistilBERT terdiri dari pembersihan data, tokenisasi, dan *fine-tuning* DistilBERT menggunakan *AdamW optimizer*. Kemudian, dilakukan eksplorasi 24 kombinasi *hyperparameter tuning*, dari beberapa nilai pada *learning rate*, *batch size*, *dropout*, dan *weight decay* untuk memperoleh konfigurasi terbaik. Model terbaik mencapai *validation accuracy* sebesar 91,86%, dengan konfigurasi nilai *hyperparameter learning rate* 0,00001, *batch size* 16, *dropout* 0,3, dan *weight decay* 0,01. Model terbaik yang diuji menghasilkan performa *accuracy* 88,44%, *precision* 88,43%, *recall* 88,53%, dan *F1-score* 88,43%. Hasil penelitian menunjukkan bahwa model DistilBERT memiliki kemampuan generalisasi yang baik dalam mengklasifikasikan komentar *cyberbullying* berbahasa Indonesia, sehingga berpotensi mendukung terciptanya ekosistem digital yang lebih aman.

Kata kunci :deteksi *cyberbullying*, DistilBERT, *natural language processing*, klasifikasi teks