

## ABSTRAK

Perkembangan media sosial, khususnya YouTube, menyebabkan meningkatnya interaksi pengguna pada kolom komentar. Tingginya aktivitas komentar tersebut juga memunculkan berbagai bentuk komentar negatif seperti penghinaan, pelecehan, sarkasme, dan serangan verbal yang termasuk dalam kategori *cyberbullying*. Oleh karena itu, diperlukan sebuah sistem yang mampu mengidentifikasi komentar *cyberbullying* secara otomatis. Penelitian ini bertujuan untuk membangun sistem deteksi *cyberbullying* pada komentar YouTube berbahasa Indonesia menggunakan model *IndoBERT*. *Dataset* penelitian diperoleh melalui proses scraping komentar YouTube menggunakan YouTube Data *API* dari video konten di Indonesia tahun 2019–2024. Tahapan penelitian meliputi *preprocessing* data, *labeling* manual, *dataset balancing*, tokenisasi, *training* model, evaluasi model, serta implementasi sistem berbasis *web* menggunakan *Laravel* dan *FastAPI*. *Dataset* akhir yang digunakan sebanyak 10.000 data dengan perbandingan 60% data *cyberbullying* dan 40% data *non-cyberbullying*. Proses *training* menggunakan model *indobenchmark/indobert-base-pl* dengan pembagian data *training* dan *testing* sebesar 80:20. Berdasarkan hasil evaluasi model, diperoleh nilai *accuracy* sebesar 88%, *precision* sebesar 88%, *recall* sebesar 88%, dan *F1-score* sebesar 88%. Hasil tersebut menunjukkan bahwa model *IndoBERT* mampu melakukan klasifikasi komentar *cyberbullying* dengan cukup baik pada komentar YouTube berbahasa Indonesia. Berdasarkan hasil penelitian, sistem yang dibangun mampu membantu proses identifikasi komentar *cyberbullying* secara otomatis serta menunjukkan bahwa model *IndoBERT* efektif digunakan dalam analisis teks media sosial berbahasa Indonesia.

**Kata Kunci:** *Cyberbullying, IndoBERT, YouTube, Natural Language Processing, Klasifikasi Teks.*