ABSTRAK

Cyberbullying merupakan salah satu permasalahan sosial yang semakin meningkat seiring dengan berkembangnya *platform* media sosial. Dampak dari *cyberbullying*, seperti *hate comments*, dapat memberikan efek negatif yang besar secara psikologis dan emosional pada korban. Penelitian ini memiliki latar belakang pentingnya pengembangan sistem deteksi *hate comments* yang akurat untuk memitigasi dampak tersebut. Tujuan penelitian ini adalah untuk menganalisis performansi model IndoBERT dan Cendol dalam mendeteksi komentar kebencian pada kasus *cyberbullying*. Metode yang digunakan dalam penelitian ini adalah pengumpulan dataset yang mengandung kata kunci yang mengarah ke *hate comments* dan *bullying*, serta pengujian model IndoBERT dan Cendol. Evaluasi dilakukan menggunakan matriks seperti akurasi, presisi, *recall*, dan *F1-Score*.

Hasil penelitian menunjukkan bahwa model IndoBERT dan Cendol memiliki performansi yang kompetitif dalam mendeteksi *hate comments*. Model IndoBERT menunjukkan keunggulan dalam menangani teks formal, sedangkan model Cendol lebih efisien dalam memahami bahasa informal dan konteks lokal. Survei yang dilakukan dengan 328 responden menghasilkan 64 kata kunci yang mengandung kata *cyberbullying*. Model IndoBERT menunjukkan akurasi tertinggi dengan nilai 90,7% pada *epoch* ke-5, *learning rate* 10^{-5}, dan *batch size* 8. Sementara itu, model Cendol memperoleh akurasi 90,6% pada *epoch* ke-5, *learning rate* 10^{-4}, dan *batch size* 2. Media sosial yang digunakan dalam penelitian ini adalah X (sebelumnya Twitter), yang menjadi salah satu platform yang banyak digunakan untuk berinteraksi dan berbagi opini.

Meskipun performansi kedua model sangat baik, terdapat kemungkinan model mengalami *overfitting*. Hal ini menjadi perhatian penting untuk penelitian selanjutnya, karena *overfitting* dapat mengurangi kemampuan model untuk generalisasi pada data yang belum terlihat. Penelitian ini memberikan kontribusi penting dalam pengembangan sistem deteksi ujaran kebencian berbasis bahasa Indonesia, yang dapat diterapkan tidak hanya di X, tetapi juga di media sosial lain seperti Facebook, Instagram, dan TikTok, untuk memitigasi dampak negatif dari *hate comments* di berbagai *platform*.

Kata Kunci: *Cyberbullying*, *Hate Comments*, IndoBERT, Cendol, NLP.