Deteksi Ujaran Kebencian pada Twitter Indonesia Menggunakan Metode IndoBERTweet

Bijak Algifan Putra[1], Yuliant Sibaroni[2], Sri Suryani Prasetyowati[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]balgifanp@student.telkomuniversity.ac.id, [2]yuliant@telkimuniversity.ac.id,
[3]srisuryani@telkomuniversity.ac.id

Abstrak

Media sosial digunakan oleh masyarat sebagai media untuk berkomunikasi secara daring. Tidak jarang *platform* ini dimaanfaatkan untuk hal yang tidak baik yaitu melakukan ujaran kebencian. Ujaran kebencian pada media sosial harus dideteksi untuk menghindari terjadinya konflik antara warga dan terhindar dari contoh buruk untuk pengguna sosial media. ini mendorong para peneliti untuk membuat sebuah sistem pendeteksi ujaran kebencian secara otomatis, pada penelitian sebelumnya dilakukan dengan menggunakan pendeketan *machine learning*. Tetapi dalam beberapa tahun ini model bahasa menggunakan *pre-trained* menunjukkan terobosan besar dalam bidang NLP (*Natural language processing*). Beberapa penelitian yang telah menggunakan metode pre-trained model seperti deteksi ujaran kebencian dan bahasa kasar pada Twitter menggunakan *Bidirectional Encoder Representations From Transformers* (BERT) dan deteksi penggunaan kalimat *abusive* pada teks Indonesia menggunakan *Indonesia Bidirectional Encoder Representations from Transformers* (IndoBERT). Pada penelitian ini digunakan metode IndoBERTweet untuk melakukan pendeteksian ujaran kebencian pada Twitter. Pembuatan sistem pengujian dan pelatihan melibatkan pembangunan *dataset* dengan *crawling data* dan pelabelan data secara manual. Model IndoBERTweet telah menunjukkan kinerja yang sangat baik dalam mendeteksi ujaran kebencian dengan hasil optimal. Hasil evaluasi menunjukkan bahwa melalui *tuning hyperparameter* untuk mendapatkan model terbaik, didapatkan model mencapai akurasi sebesar 87.1%.

Kata kunci: ujaran kebencian, indobertweet, pre-trained, hyperparameter