INTISARI

DETEKSI UJARAN KEBENCIAN DAN BAHASA OFENSIF MENGGUNAKAN *BERT MODEL* DAN *DEEP LEARNING ENSAMBLE*

oleh

Fadila Shely Amalia

22/501751/PPA/06396

Deteksi ujaran kebencian adalah isu krusial dalam analisis sentimen dan pemrosesan bahasa alami. Penelitian ini bertujuan untuk meningkatkan efektivitas deteksi ujaran kebencian dalam teks berbahasa Inggris menggunakan model *Bidirectional Encoder Representations from Transformers* (BERT). Kami mengembangkan metode *preprocessing* yang dimodifikasi untuk meningkatkan nilai *F1-score*. Dataset yang digunakan terdiri dari teks berbahasa Inggris yang mengandung konten ujaran kebencian. Hasil evaluasi menunjukkan peningkatan signifikan dalam akurasi dan kinerja keseluruhan model dalam tugas deteksi teks. Model BERT mencapai akurasi sebesar 89.11% pada data uji, yang menunjukkan kemampuannya untuk memprediksi dengan benar sekitar 85 dari 95 sampel. Analisis *confusion matrix* mengindikasikan bahwa model ini sangat efektif dalam mendeteksi teks ofensif dengan akurasi sekitar 95%, meskipun mengalami kesulitan dalam membedakan antara teks kebencian dan ofensif, serta terdapat kebingungan antara teks netral dan ofensif. Dari hasil classification report, diperoleh *F1-score* sebesar 0,43 untuk kelas kebencian, 0.94 untuk kelas ofensif, dan 0.84 untuk kelas netral. *Weighted average F1-score* mencapai 0.89. Penelitian ini menunjukkan bahwa meskipun penggabungan BERT dengan metode *deep learning* lain seperti CNN dan DNN telah diuji, BERT tetap unggul dengan *weighted F1-score* 0.89, lebih tinggi dibandingkan BERT + CNN (0.85) dan BERT + DNN (0.67). Dengan demikian, BERT diandalkan sebagai model terbaik dalam deteksi konten berbasis teks.

Kata Kunci: *Hate speech*, *Offensive*, *Deep Learning*, BERT, Twitter

ABSTRACT

HATE SPEECH AND OFFENSIVE LANGUAGE DETECTION USING BERT MODEL AND DEEP LEARNING ENSAMBLE

by

Fadila Shely Amalia

22/501751/PPA/06396

Hate speech detection is a crucial issue in sentiment analysis and natural language processing. This study aims to enhance the effectiveness of hate speech detection in English text using the Bidirectional Encoder Representations from Transformers (BERT) model. We developed a modified preprocessing method to improve the F1-score. The dataset used consists of English text containing hate speech content. Evaluation results show a significant increase in the accuracy and overall performance of the model in the text detection task. The BERT model achieved an accuracy of 89.11% on the test data, indicating its ability to correctly predict approximately 85 out of 95 samples. Confusion matrix analysis indicates that the model is highly effective in detecting offensive text, with an accuracy of around 95%, although it struggles to differentiate between hate speech and offensive text, and there is some confusion between neutral and offensive text. From the classification report, an F1-score of 0.43 was obtained for the hate class, 0.94 for the offensive class, and 0.84 for the neutral class. The weighted average F1-score reached 0.89, while the macro average F1-score was 0.73. This research demonstrates that despite the tested combinations of BERT with other deep learning methods such as CNN and DNN, BERT remains superior with a weighted F1-score of 0.89, higher than BERT + CNN (0.85) and BERT + DNN (0.67). Thus, BERT is relied upon as the best model for detecting text-based content that requires a complex understanding of language context, showing significant improvement compared to previous research, which only achieved a weighted F1-score of 0.762.

***Keywords*: *Hate speech*, *Offensive*, *Deep Learning*, BERT, Twitter**