INTISARI

DETEKSI UJARAN KEBENCIAN PADA TEKS BAHASA INGGRIS MENGGUNAKAN TRANSFER LEARNING HATEBERT DAN PENERAPAN EXPLAINABLE AI BERUPA LIME

Oleh

Konang Tyagazain Nirangkara

21/474140/PA/20469

Keterbatasan metode berbasis kata kunci dan model machine learning tradisional dalam mendeteksi ujaran kebencian yang bersifat tersirat, sarkasme, atau istilah terselubung sering kali mengurangi akurasi sistem deteksi. Penelitian ini mengusulkan penerapan *transfer learning* pada HateBERT, model berbasis *transformer* yang telah dilatih dengan dataset ujaran kebencian, untuk meningkatkan pemahaman konteks dalam klasifikasi teks pada dataset OffensEval 2019. Dengan menerapkan pra-pemrosesan teks, tokenisasi, dan *fine-tuning* model, pendekatan ini memungkinkan model menangkap makna ujaran kebencian secara lebih efektif. Selain itu, penelitian ini mengintegrasikan Explainable AI (XAI) menggunakan LIME untuk menganalisis kontribusi setiap kata terhadap keputusan model, sehingga meningkatkan transparansi dan interpretabilitas model deteksi ujaran kebencian. Evaluasi dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, *macro F1-score*, dan *POS F1-score*, serta analisis interpretabilitas model melalui hasil XAI berupa LIME. Setelah dilakukan *transfer learning*, model mencapai *macro F1-score* sebesar 0.7878, menunjukkan performa yang baik dalam klasifikasi ujaran kebencian. Implementasi LIME menghasilkan visualisasi kontribusi kata berupa *highlight* warna, yang membantu pengguna memahami faktor-faktor yang memengaruhi prediksi model. Namun, penjelasan yang diberikan bersifat lokal dan tidak sepenuhnya konsisten, sehingga tetap diperlukan peran manusia untuk melakukan interpretasi yang tepat atas hasil yang ditampilkan.

Kata Kunci: Deteksi ujaran kebencian, HateBERT, Transfer learning, Explainable AI, LIME, NLP

ABSTRACT

HATE SPEECH DETECTION IN ENGLISH TEXT USING HATEBERT TRANSFER LEARNING AND EXPLAINABLE AI IMPLEMENTATION USING LIME

By

Konang Tyagazain Nirangkara

21/474140/PA/20469

The limitations of keyword-based methods and traditional machine learning models in detecting implied hate speech, sarcasm, or veiled terms often reduce the accuracy of detection systems. This research proposes applying transfer learning to HateBERT, a transformer-based model that has been trained with hate speech datasets, to improve context understanding in text classification on the OffensEval 2019 dataset. By applying text pre-processing, tokenization, and model fine-tuning, this approach enables the model to capture the meaning of hate speech more effectively. In addition, this research integrates Explainable AI (XAI) using LIME to analyze the contribution of each word to the model's decision, thus improving the transparency and interpretability of the hate speech detection model. Evaluation was conducted using accuracy, precision, recall, macro F1-score, and POS F1-score metrics, as well as model interpretability analysis through XAI results in the form of LIME. After transfer learning, the model achieved a macro F1-score of 0.7878, indicating good performance in hate speech classification. The LIME implementation produces a visualization of word contributions in the form of color highlights, which helps users understand the factors that influence the model's predictions. However, the explanations provided are localized and not fully consistent, so a human is still required to make a proper interpretation of the results.

Keywords: Hate speech detection, HateBERT, Transfer learning, Explainable AI, LIME, NLP