

COMPARATIVE ANALYSIS OF HATE SPEECH CLASSIFICATION MODELS ON SOCIAL MEDIA: BERT, XGBOOST, AND HYBRID BERT-XGBOOST

ABSTRAK

Perkembangan media sosial telah memicu peningkatan signifikan dalam penyebaran ujaran kebencian daring, khususnya pada platform X (Twitter). Penelitian ini mengusulkan pendekatan hybrid berbasis Bidirectional Encoder Representations from Transformers (BERT) dan algoritma Extreme Gradient Boosting (XGBoost) untuk mendeteksi konten bermuatan ujaran kebencian secara lebih akurat, sekaligus mengintegrasikan Generative AI untuk menghasilkan penjelasan naratif yang mudah dipahami oleh audiens non-teknis. Dataset dikumpulkan melalui scraping menggunakan Twitter Search API dan anotasi berbasis crowdsourcing, mencakup kategori, target, dan intensitas ujaran kebencian, dengan total sebanyak 13.014 data. Tiga model dikembangkan dan dibandingkan: BERT + XGBoost, XGBoost berbasis TF-IDF, dan BERT finetuned. Hasil evaluasi menunjukkan bahwa model hybrid mencapai akurasi 81%, namun tidak melampaui performa model BERT finetuned, yang memperoleh akurasi 88,99% dan F1-score 0,8893. Temuan ini mengindikasikan bahwa pendekatan hybrid tidak selalu menjamin peningkatan kinerja dalam konteks klasifikasi teks berbahasa Indonesia.

PENDAHULUAN

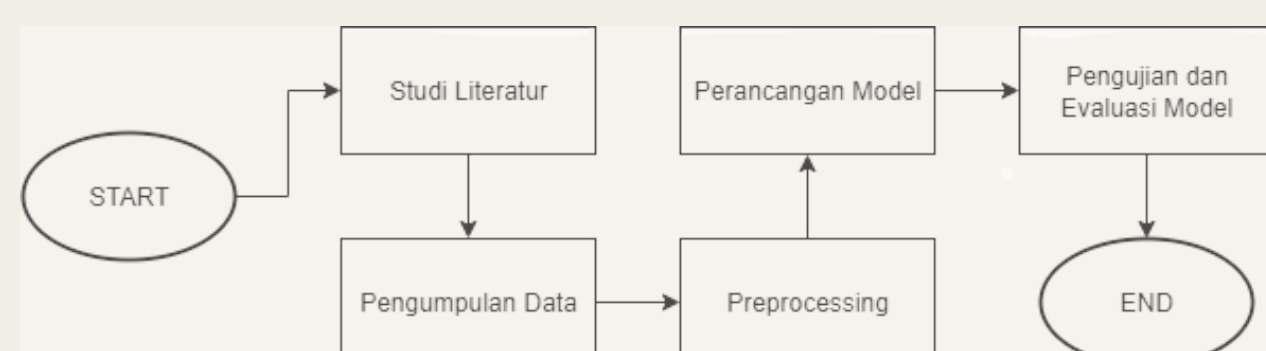
Perkembangan teknologi digital telah memicu perubahan besar dalam berbagai aspek kehidupan, termasuk meningkatnya penyebaran ujaran kebencian di ruang digital. Ujaran kebencian merupakan ekspresi yang menyerang individu atau kelompok berdasarkan atribut seperti ras, agama, atau etnis, dan sering kali memanfaatkan anonimitas di media sosial [1]. Dampaknya mencakup gangguan psikologis, polarisasi sosial, hingga potensi konflik di dunia nyata.

Platform X (Twitter) menjadi salah satu media utama penyebaran ujaran kebencian di Indonesia [2]. Fitur seperti retweet dan hashtag mempercepat penyebaran konten agresif, terutama yang menargetkan kelompok minoritas. Laporan terbaru menunjukkan peningkatan sepuluh kali lipat dalam dua tahun terakhir, menimbulkan kekhawatiran terhadap keamanan dan kualitas interaksi di ruang digital [3].

Pendekatan Model BERT (Bidirectional Encoder Representations from Transformers) merupakan perkembangan revolusioner dengan kemampuannya dalam memahami konteks semantik [4]. Studi sebelumnya menunjukkan performa tinggi BERT dalam deteksi ujaran kebencian, seperti F1-score 92% untuk bahasa Inggris dan macro-F1 0,78 untuk bahasa Indonesia [5, 2].

Penelitian ini mengusulkan pendekatan hybrid dengan menggabungkan representasi semantik dari BERT dan kemampuan klasifikasi tinggi dari XGBoost (Amalia & Suryani, 2021; Babaeianjelodar et al., 2022). Pendekatan ini diharapkan dapat meningkatkan akurasi deteksi ujaran kebencian di media sosial Indonesia secara lebih efektif.

METODOLOGI PENELITIAN



Skenario	Sentimen	Akurasi	Metrics (%)		
			Precision	Recall	F1-Score
BERT + XGBoost	Hate-Speech	81%	81%	81%	81%
	Non-Hate		81%	81%	91%
	Avg.		81%	81%	81%
XGBoost TF-IDF	Hate-Speech	85%	83%	89%	86%
	Non-Hate		88%	83%	85%
	Avg.		86%	86%	86%
BERT Finetuned	Hate-Speech	90%	89%	90%	89%
	Non-Hate		90%	89%	89%
	Avg.		90%	90%	89%

HASIL PENELITIAN

Model BERT Finetuned menunjukkan kinerja terbaik secara keseluruhan dengan akurasi sebesar 90% dan skor evaluasi rata-rata yang seimbang pada semua metrik (precision, recall, dan F1-score sebesar 89–90%). Model XGBoost TF-IDF menempati posisi kedua setelah BERT finetuned dengan akurasi 85% dan skor metrik yang cukup kompetitif, khususnya pada kelas Hate Speech yang menunjukkan recall sebesar 89%. Sementara itu, model BERT + XGBoost menunjukkan performa paling rendah di antara ketiganya, dengan akurasi hanya 81% dan skor metrik yang relatif seragam namun lebih rendah.

Hasil ini menunjukkan bahwa penggabungan representasi vektor embeddings dari BERT dengan model pembelajaran klasik seperti XGBoost tidak secara otomatis menghasilkan peningkatan performa. Salah satu penyebabnya adalah kurang optimalnya proses integrasi representasi vektor kontekstual dengan arsitektur model pembelajaran klasik yang tidak dirancang untuk memanfaatkan konteks semantik mendalam sebagaimana halnya jaringan neural transformer. Selain itu, pendekatan ini juga tidak melibatkan pelatihan ulang parameter BERT, sehingga potensi penuh dari representasi bahasa yang kontekstual tidak dapat dimanfaatkan secara maksimal.

KESIMPULAN

Penelitian ini mengimplementasikan dan membandingkan tiga model klasifikasi ujaran kebencian pada platform X, yaitu model hybrid BERT + XGBoost, XGBoost berbasis fitur TF-IDF, dan BERT finetuned. Hasil evaluasi menunjukkan bahwa model BERT finetuned memberikan performa terbaik dengan akurasi 88,99%, precision 0,89, recall 0,89, dan F1-score 0,89. Model hybrid BERT + XGBoost mencapai akurasi 81% namun tidak mampu mengungguli model BERT finetuned maupun XGBoost TF-IDF. Keunggulan model BERT finetuned terletak pada kemampuannya memahami konteks linguistik dan nuansa semantik ujaran kebencian bahasa Indonesia secara efektif. Namun, tantangan seperti sarkasme, idiom, dan konteks sosial budaya masih menjadi hambatan yang memerlukan pengembangan lebih lanjut untuk meningkatkan sensitivitas model terhadap variasi bahasa alami.

[1] C.-H. Lee and M. Sanchez, "Cyberbullying: Prevalence, Causes, and Consequences," *International Journal of Cyber Criminology*, vol. 12, no. 1, pp. 78–95, 2018.

[2] I. Alfina *et al*, "IndoToxic2024: Dataset for Hate Speech in Indonesian," *arXiv preprint* arXiv:2406.19349*, 2024.

[3] G. Ray, C. D. McDermott, and M. Nicho, "Cyberbullying on Social Media: Definitions, Prevalence, and Impact Challenges," *Journal of Cybersecurity*, vol. 10, no. 1, p. tyae026, 2024.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers," *arXiv preprint* arXiv:1810.04805*, 2019.

[5] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Hate speech detection and racial bias mitigation," *PLOS ONE*, vol. 15, no. 8, p. e0237861, 2020, doi: 10.1371/journal.pone.0237861.

[6] F. S. Amalia and Y. Suyanto, "Offensive language and hate speech detection using BERT model," *Indonesian Journal of Computing and Cybernetics Systems*, vol. 15, no. 2, pp. 129–136, 2021, doi: 10.22146/ijccs.99841.