KLASIFIKASI UJARAN KEBENCIAN PADA MEDIA SOSIAL INSTAGRAM MENGGUNAKAN PENDEKATAN TRANSFER LEARNING DENGAN FINE-TUNING INDOBERT

Panji Angkasa Putra[1)], Rachmatu Zabila Khaerunissa[2)], Rahma Nursapitri[3)], Rafli Indra Gunawan[4)]
[1, 2, 3, 4)]Teknik Informatika, Fakultas Teknik Komputer dan Desain, Universitas Nusa Putra
Jl. Raya Cibolang No.21, Cibolang Kaler, Kec.Cisaat, Kabupaten Sukabumi, Jawa Barat 43152
e-mail: panji.angkasa_ti23@nusaputra.ac.id[1)], rachmatu.zabila_ti23@nusaputra.ac.id[2)], rahma.nursapitri_ti23@nusaputra.ac.id[3)], rafli.indra_ti23@nusaputra.ac.id[4)]

* Korespondensi: e-mail: panji.angkasa_ti23@nusaputra.ac.id

ABSTRAK

Media sosial Instagram telah berkembang menjadi salah satu platform interaksi digital terbesar di Indonesia. Namun, kebebasan berinteraksi ini sering disalahgunakan untuk menyebarkan ujaran kebencian yang dapat memicu konflik sosial dan dampak psikologis negatif. Tantangan utama dalam mendeteksi ujaran kebencian secara otomatis adalah kompleksitas bahasa Indonesia yang tidak baku, penggunaan singkatan, serta konteks kalimat yang ambigu. Penelitian ini mengusulkan pendekatan Transfer Learning menggunakan model bahasa pre-trained IndoBERT (Indonesian Bidirectional Encoder Representations from Transformers) yang di-fine-tune untuk mengklasifikasikan komentar Instagram ke dalam dua kelas: ujaran kebencian dan bukan ujaran kebencian. Dataset yang digunakan terdiri dari 14.306 data. Proses evaluasi model dilakukan menggunakan Confusion Matrix dan kurva ROC-AUC (Receiver Operating Characteristic - Area Under the Curve). Hasil penelitian menunjukkan bahwa model IndoBERT mampu mengenali pola bahasa kompleks dengan performa yang baik, mencapai Akurasi sebesar 90% dan F1-Score sebesar 90%. Selain itu, nilai AUC sebesar 0.9504 menunjukkan kemampuan model yang sangat baik dalam memisahkan kelas ujaran kebencian atau bukan. Penelitian ini membuktikan bahwa teknik fine-tuning pada model transformer berbahasa Indonesia efektif untuk memitigasi penyebaran komentar toksik di platform Instagram.

***Kata Kunci**: Ujaran Kebencian, Instagram, Klasifikasi Teks, Transfer Learning, IndoBERT*

ABSTRACT

Instagram has become one of the largest digital interaction platforms in Indonesia. However, this freedom of interaction is often misused to spread hate speech, which can trigger social conflict and negative psychological impacts. The main challenges in automatically detecting hate speech are the complexity of non-standard Indonesian, the use of abbreviations, and ambiguous sentence contexts. This study proposes a Transfer Learning approach using a pre-trained language model, IndoBERT (Indonesian Bidirectional Encoder Representations from Transformers), which has been fine-tuned to classify Instagram comments into two classes: hate speech and non-hate speech. The dataset used consists of 14,306 data points. The model evaluation process was carried out using a Confusion Matrix and ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) curve. The results show that the IndoBERT model is able to recognize complex language patterns with good performance, achieving 90% accuracy and 90% F1-Score. In addition, the AUC value of 0.9504 indicates the model's excellent ability to distinguish hate speech from non-hate speech. This study proves that the fine-tuning technique on the Indonesian language transformer model is effective in mitigating the spread of toxic comments on the Instagram platform.

***Keywords:** Hate Speech, Instagram, Text Classification, Transfer Learning, IndoBERT*

I. PENDAHULUAN

A. *Latar Belakang*

Perkembangan teknologi informasi telah merubah sarana komunikasi masyarakat Indonesia. Model komunikasi kewargaan yang selama ini bersifat konvensional, perlahan bertransformasi menjadi pola komunikasi berbasis internet dan teknologi digital. Pola komunikasi kewargaan semakin tampak lebih taktis, cepat, efektif dan efisien [1]. Salah satu platform sosial media yang memiliki pengguna aktif jutaan di Indonesia yaitu Instagram.

Instagram tidak hanya memiliki fitur untuk memposting sebuah foto atau video saja, akan tetapi ada juga fitur komentar. Fitur ini digunakan saat konsumen atau followers mengomentari sebuah unggahan [2]. Dalam hal ini pengguna memiliki kebebasan untuk memberikan sebuah respon terhadap sebuah postingan, tetapi dengan kebebasan tersebut banyak juga yang memberikan respon ujaran kebencian.

Ujaran kebencian adalah tindakan komunikasi yang dilakukan oleh suatu individu atau kelompok dalam bentuk provokasi, hasutan, ataupun hinaan kepada individu atau kelompok yang lain dalam hal berbagai aspek seperti ras, warna kulit, etnis, gender, cacat, orientasi seksual, kewarganegaraan, agama, dan lain-lain [3]. Perilaku tersebut jika dibiarkan, maka akan menyebarkan dampak negatif yang lebih luas. Maka dari itu dibutuhkan sistem berbasis kecerdasan buatan (Artificial Intelligence) untuk mengatasi masalah tersebut.

Teknologi kecerdasan buatan dapat menjadi solusi untuk mengidentifikasi ujaran kebencian pada komentar di media sosial Instagram. Salah satu metode yang efektif digunakan dalam pemrosesan teks adalah pendekatan Transfer Learning dengan teknik Fine-Tuning. Fine-tuning mengacu pada proses mengambil model neural network yang telah dilatih sebelumnya dan menyesuaikannya untuk tujuan spesifik tugas baru [4].

Penelitian sebelumnya telah mengkaji deteksi ujaran kebencian pada media sosial menggunakan pendekatan machine learning. Salah satu penelitian menerapkan metode Naïve Bayes dengan pendekatan N-Gram pada dataset Twitter berbahasa Indonesia untuk mengklasifikasikan ujaran kebencian beserta tingkat ancamannya dan memperoleh nilai F-score sebesar 64,957%, yang menunjukkan keterbatasan metode probabilistik dalam menangkap kompleksitas konteks bahasa [5]. Penelitian lain menggunakan algoritma K-Nearest Neighbor (KNN) untuk mendeteksi sentimen ujaran kebencian pada tweet di platform media sosial X dalam konteks RKUHP, dengan hasil yang menunjukkan performa deteksi yang moderat, namun masih menghadapi kendala dalam menangani sarkasme, ambiguitas makna, dan netralitas sentimen [6].

Berdasarkan keterbatasan metode klasifikasi konvensional dalam memahami kompleksitas bahasa Indonesia, diperlukan pendekatan yang mampu menangkap konteks semantik secara lebih mendalam. Oleh karena itu, penelitian ini mengusulkan penggunaan pendekatan Transfer Learning dengan melakukan fine-tuning pada model bahasa pre-trained IndoBERT untuk mengklasifikasikan komentar Instagram ke dalam dua kelas, yaitu ujaran kebencian dan bukan ujaran kebencian. Pendekatan ini diharapkan mampu meningkatkan akurasi deteksi ujaran kebencian serta berkontribusi dalam mitigasi penyebaran komentar toksik di platform media sosial Instagram.

II. KAJIAN TEORI

A. *Ujaran Kebencian*

Ujaran kebencian yaitu kegiatan seseorang melalui perkataan, perbuatan, tulisan maupun pertunjukan dengan maksud untuk menghina, memprovokasi, ataupun menghasut orang lain dengan tujuan untuk membuat prasangka baik ditunjukkan untuk pelaku ujaran kebencian tersebut maupun korban dari tindakan itu sendiri [7]. Definisi ini menegaskan bahwa ujaran kebencian tidak hanya terbatas pada bentuk verbal, tetapi juga mencakup tindakan nonverbal yang berpotensi memicu konflik sosial dan merugikan pihak tertentu. Ujaran kebencian dapat menimbulkan dampak negatif, seperti diskriminasi dan gangguan psikologis bagi korban.

B. *Natural Language Processing (NLP)*

Natural Language Processing (NLP) adalah salah satu bidang ilmu komputer yang merupakan

cabang dari kecerdasan buatan, dan bahasa (linguistik) yang berkaitan dengan interaksi antara komputer dan bahasa alami manusia, seperti bahasa Indonesia atau bahasa Inggris [8]. Pendekatan NLP memungkinkan sistem komputer untuk memahami, menganalisis, dan menghasilkan bahasa manusia secara otomatis, sehingga sangat relevan dalam pengolahan teks pada media sosial. Dalam hal ini, jika diterapkan komputer dapat memberikan respon bahasa yang lebih natural mirip dengan bahasa alami manusia.

C. *Transfer Learning*

Transfer learning adalah suatu metode pembelajaran yang mendalam di mana pengetahuan yang diperoleh dari mempelajari satu masalah digunakan untuk memecahkan masalah yang berbeda-beda [9]. Konsep ini memungkinkan model memanfaatkan representasi pengetahuan yang telah dipelajari sebelumnya sehingga dapat meningkatkan kinerja dan efisiensi pelatihan pada tugas baru dengan jumlah data terbatas.

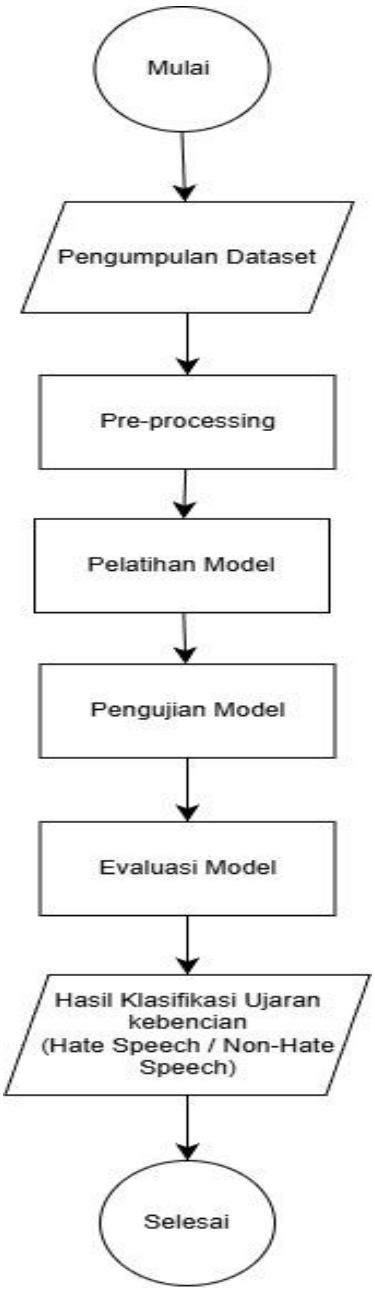
D. *IndoBERT (Indonesian BERT)*

BERT menggunakan Transformer, sebuah mekanisme yang mempelajari hubungan kontekstual antara kata-kata dalam teks dengan menggunakan self-attention mechanism. Khusus untuk bahasa Indonesia pada tahun 2020, berhasil dikembangkan model pre-trained BERT yang disebut IndoBERT [10]. IndoBERT dirancang untuk menangkap karakteristik linguistik bahasa Indonesia, sehingga lebih efektif digunakan dalam berbagai tugas pemrosesan bahasa alami berbahasa Indonesia.

E. *Fine-Tuning*

Fine-tuning adalah proses mengadaptasi pengetahuan umum dari model pre-trained untuk dispesialisasi dalam tugas tertentu yang dalam hal ini adalah klasifikasi emosi dengan melatihnya pada dataset yang relevan [11]. Melalui proses fine-tuning, model dapat menyesuaikan bobot parameternya agar lebih optimal dalam mengenali pola spesifik sesuai dengan tujuan penelitian.

III. METODE



Gambar I. Diagram Alur Sistem
(Sumber: Olahan Peneliti)

A. Pengumpulan Dataset

Pada tahap ini, dilakukan pengumpulan data ujaran kebencian dengan dua kelas yang diambil dari platform HuggingFace. Total dari dataset adalah 14306 data, berikut tabel dari datasetnya:

Tabel I. Jumlah Dataset
(Sumber: Dataset HuggingFace [Indonesian-Hate-Speech-Superset], diolah peneliti)

Kelas	Jumlah Data
Ujaran Kebencian	6050
Bukan Ujaran Kebencian	8256

B. Preprocessing

Tahap pra-pemrosesan data bertujuan untuk mengubah data teks mentah menjadi format terstruktur yang dapat dipahami oleh model Deep Learning. Mengingat data yang digunakan berasal dari media sosial yang memiliki karakteristik tidak terstruktur, tahapan ini sangat krusial

untuk memastikan kualitas input model. Pada penelitian ini, tahapan pra-pemrosesan meliputi tiga proses utama: filtrasi data, pembersihan data, dan tokenisasi.

C. *Training*

Penelitian ini menerapkan metode fine-tuning pada model IndoBERT-base-p1 dengan menambahkan lapisan klasifikasi linear pada output token. Lapisan ini bertugas memetakan representasi fitur dari IndoBERT ke dalam dua kelas output (Ujaran Kebencian dan Non-Ujaran Kebencian) dengan meminimalkan fungsi kerugian Cross-Entropy Loss.

Proses pelatihan dilakukan menggunakan framework PyTorch dan Hugging Face dengan konfigurasi hiperparameter sebagai berikut:

Tabel II. Parameter pelatihan IndoBERT
(Sumber: Olahan Peneliti)

Nama	Parameter
Optimizer	AdamW
Learning Rate	2e-5
Epoch	3
Batch Size	4

Model dievaluasi pada setiap akhir epoch, dan pos-el (checkpoint) model dengan nilai F1-Score terbaik akan disimpan sebagai model final.

D. *Pengujian Model*

Pengujian model dilakukan menggunakan data uji sebesar 20% dari total dataset yang dipisahkan sebelum proses pelatihan. Langkah ini bertujuan untuk mengukur kemampuan generalisasi model dalam mengklasifikasikan data baru yang belum pernah dipelajari sebelumnya.

E. *Evaluasi Model*

Evaluasi kinerja didasarkan pada metrik Akurasi untuk mengukur ketepatan prediksi secara global, serta F1-Score sebagai metrik utama. Penggunaan F1-Score sangat krusial mengingat karakteristik dataset ujaran kebencian yang sering kali tidak seimbang, sehingga rata-rata harmonik antara Precision dan Recall memberikan gambaran performa yang lebih objektif. Selain itu, analisis kesalahan dilakukan menggunakan Confusion Matrix untuk memetakan tingkat False Positive dan False Negative, serta kurva ROC-AUC untuk mengukur kemampuan model dalam membedakan antar kelas pada berbagai ambang batas. Berikut adalah rumus pengukuran evaluasi performance pada confusion matrix:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F1-Score} = \frac{2\times precision\times recall}{precision+recall} \quad (4)$$

Keterangan:

- Accuracy : Rasio antara jumlah prediksi benar (baik positif maupun negatif) dengan total jumlah prediksi yang dibuat.
- Precision : Rasio antara jumlah prediksi positif yang benar dengan total prediksi positif yang dibuat oleh model.
- Recall : Rasio antara jumlah prediksi positif yang benar dengan total jumlah data yang sebenarnya positif.
- F1-Score : Nilai yang menggabungkan precision dan recall untuk memberikan gambaran

keseluruhan tentang kinerja model.

IV. HASIL DAN PEMBAHASAN

A. *Evaluasi Kinerja Model Berdasarkan Confusion Matrix*

Penelitian ini berfokus pada penerapan metode Transfer Learning dengan melakukan Fine-Tuning pada model pre-trained IndoBERT-base-p1. Tujuan utama dari rangkaian eksperimen ini adalah untuk mengukur sejauh mana kemampuan model dalam mengenali dan membedakan pola linguistik antara komentar yang mengandung ujaran kebencian (hate speech) dan komentar yang tidak mengandung ujaran kebencian (non-hate speech) pada platform Instagram.

Proses pengujian dilakukan menggunakan data uji (test set) yang telah dipisahkan sebesar 20% dari total dataset. Data ini merupakan data baru yang belum pernah dilihat oleh model selama proses pelatihan (training), sehingga sangat representatif untuk mengukur kemampuan generalisasi model di dunia nyata. Indikator keberhasilan model dievaluasi secara komprehensif menggunakan berbagai metrik standar, meliputi Confusion Matrix untuk memetakan distribusi kesalahan prediksi, laporan klasifikasi (Classification Report) yang mencakup Precision, Recall, dan F1-Score, serta analisis kurva ROC-AUC untuk melihat tingkat diskriminasi model. Berikut adalah rincian hasil evaluasi kinerja model yang diperoleh.

Tabel III. Hasil Confusion Matrix
(Sumber: Olahan Peneliti)

		Kelas Aktual	
		Kelas_0	*Kelas_1*
Kelas Prediksi	*Kelas_0*	*47*	*6*
	Kelas_1	*6*	*56*

Evaluasi pertama dilakukan dengan menganalisis Confusion Matrix untuk melihat detail prediksi benar dan salah yang dihasilkan oleh model IndoBERT setelah proses pelatihan. Berdasarkan hasil pengujian pada data uji (test set), diperoleh distribusi prediksi sebagai berikut:

- True Negative (TN) : Model berhasil memprediksi 47 data komentar bersih (Kelas 0) dengan benar sebagai bukan ujaran kebencian.
- True Positive (TP) : Model berhasil mendeteksi 56 data ujaran kebencian (Kelas 1) dengan tepat.
- False Positive (FP) : Terdapat 6 data yang sebenarnya bukan ujaran kebencian, namun salah diprediksi oleh model sebagai ujaran kebencian.
- False Negative (FN) : Terdapat 6 data ujaran kebencian yang gagal dideteksi dan diprediksi sebagai komentar aman.

Hasil ini menunjukkan keseimbangan yang baik pada model, di mana tingkat kesalahan (error rate) untuk False Positive dan False Negative memiliki jumlah yang sama, yaitu 6 kasus. Hal ini mengindikasikan bahwa model tidak mengalami bias yang signifikan terhadap salah satu kelas.

B. *Evaluasi Kinerja Model Berdasarkan Classification Report*

Untuk mengukur kualitas klasifikasi secara menyeluruh, dihitung nilai Precision, Recall, dan F1-Score berdasarkan laporan klasifikasi (Classification Report).

Tabel IV. Laporan Klasifikasi
(Sumber: Olahan Peneliti)

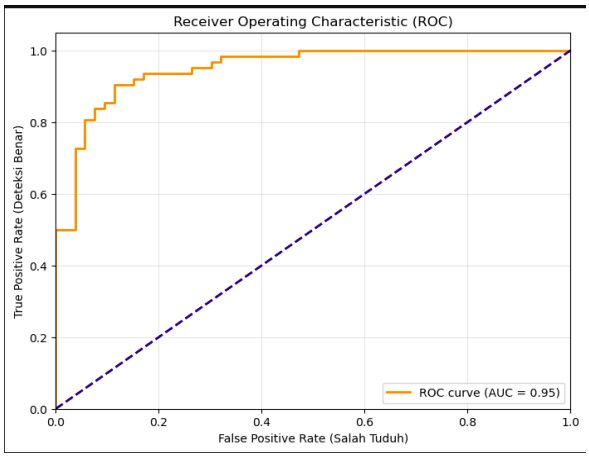
Kelas	*Precision*	*Recall*	*F1-score*
Kelas_0	*0.89*	*0.89*	*0.89*
Kelas_1	*0.90*	*0.90*	*0.90*
	Accuracy		***0.90***

- Akurasi Total : Model mencapai tingkat akurasi sebesar 0.90 atau 90% pada total 115 data uji.
- Kelas 0 : model mencatatkan precision sebesar 0.89, recall 0.89, dan f1-score 0.89.
- Kelas 1 : performa sedikit lebih tinggi dengan precision 0.90, recall 0.90, dan f1-score 0.90.

Tingginya nilai F1-Score (0.90) menegaskan bahwa model IndoBERT yang telah di-fine-tune mampu menangani tugas klasifikasi dengan sangat baik, menyeimbangkan kemampuan untuk menangkap ujaran kebencian.

C. Analisis Kurva ROC dan Nilai AUC

Selain metrik diskrit, performa model juga dievaluasi menggunakan kurva Receiver Operating Characteristic (ROC). Kurva ini menggambarkan trade-off antara True Positive Rate (TPR) dan False Positive Rate (FPR) pada berbagai ambang batas (threshold).

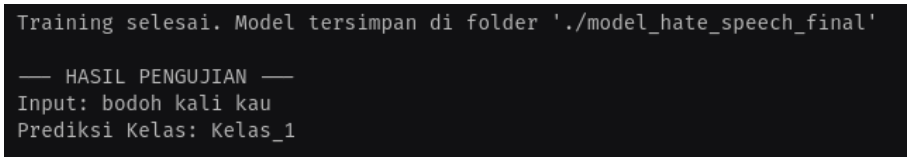


Gambar II. Kurva ROC-AUC
(Sumber: Olahan Peneliti)

Berdasarkan grafik ROC yang dihasilkan, model mencatatkan nilai AUC (Area Under Curve) sebesar 0.95. Nilai AUC yang mendekati angka 1.0 ini menunjukkan bahwa model memiliki kemampuan diskriminatif yang sangat baik dalam memisahkan antara kelas ujaran kebencian dan bukan ujaran kebencian. Kurva yang melengkung tajam ke sudut kiri atas menandakan sensitivitas yang tinggi dengan tingkat kesalahan positif palsu yang rendah.

D. Pengujian

Untuk memvalidasi model, dilakukan uji coba prediksi menggunakan input teks baru yang tidak ada dalam dataset latih.



Gambar III. Hasil Pengujian
(Sumber: Olahan Peneliti)

Hasil pengujian, model berhasil mengklasifikasikannya ke dalam Kelas_1 (Ujaran Kebencian). Hal ini membuktikan bahwa model mampu memahami konteks linguistik kata "bodoh" sebagai bentuk penghinaan atau ujaran kebencian, bukan sekadar kata sifat biasa.

V. KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan yang telah dilakukan, dapat disimpulkan bahwa penerapan metode Transfer Learning melalui teknik Fine-Tuning pada model pre-trained IndoBERT terbukti sangat efektif dalam mengklasifikasikan ujaran kebencian pada media sosial Instagram. Pendekatan ini berhasil mengatasi tantangan karakteristik linguistik teks media sosial yang cenderung tidak baku, memiliki struktur kalimat yang kompleks, serta sarat akan konteks semantik, tanpa memerlukan proses rekayasa fitur manual yang rumit. Model mampu mempelajari representasi bahasa Indonesia dengan baik sehingga dapat mengenali pola ujaran kebencian secara akurat.

Kinerja model yang dikembangkan menunjukkan hasil yang sangat signifikan pada pengujian menggunakan data uji (test set). Hal ini dibuktikan dengan perolehan nilai akurasi sebesar 90% dan F1-Score sebesar 0.90, yang mengindikasikan bahwa model memiliki kemampuan yang seimbang dan stabil, baik dalam mendeteksi komentar yang mengandung ujaran kebencian maupun dalam mengenali komentar yang aman. Selain itu, nilai Area Under Curve (AUC) yang mencapai 0.95 menegaskan bahwa model memiliki tingkat diskriminasi yang sangat tinggi dengan probabilitas kesalahan prediksi yang minim. Keberhasilan validasi pada pengujian manual semakin memperkuat kesimpulan bahwa model IndoBERT yang telah di-fine-tune ini memiliki performa yang handal dan siap diimplementasikan sebagai solusi otomatis untuk memitigasi penyebaran konten negatif di platform Instagram.

VI. DAFTAR PUSTAKA

[1] B. Arianto and B. Handayani, "Media sosial sebagai saluran komunikasi digital kewargaan: Studi etnografi digital," Arkana: Jurnal Komunikasi dan Media, vol. 2, no. 2, pp. 220–236, 2023.

[2] C. Antasari and R. D. Pratiwi, "Pemanfaatan fitur Instagram sebagai sarana komunikasi pemasaran kedai Babakkeroyokan di Kota Palu," Jurnal Ilmu Komunikasi, vol. 10, no. 1, 2022.

[3] S. Mawarti, "Fenomena hate speech," Toleransi: Media Komunikasi Umat Beragama, vol. 10, no. 1, pp. 83–95, 2018.

[4] D. Nuryadi, F. Metandi, N. A. Hadiwijaya, and M. I. U. Haq, "Fine tuning IndoBERT untuk analisis sentimen pada ulasan pengguna aplikasi Tiket.com di Google Play Store," JATI (Jurnal Mahasiswa Teknik Informatika), vol. 9, no. 2, pp. 3577–3583, 2025.

[5] R. M. Yazid, F. R. Umbara, and P. N. Sabrina, "Deteksi ujaran kebencian dengan metode klasifikasi Naïve Bayes dan metode N-Gram pada dataset multi-label Twitter berbahasa Indonesia," Informatics and Digital Expert (INDEX), vol. 4, no. 2, pp. 46–52, 2022.

[6] Y. Septiawan, J. A. Putra, A. Aglasia, D. A. Muktiawan, and Meiliza, "Deteksi sentimen ujaran kebencian dalam tweet media sosial X dalam konteks RKUHP menggunakan algoritma KNN," Jurnal Tecnoscienza, vol. 10, no. 1, pp. 54–68, 2025.

[7] F. I. Febriansyah and H. S. Purwinarto, "Pertanggungjawaban pidana bagi pelaku ujaran kebencian di media sosial," Jurnal Penelitian Hukum De Jure, vol. 20, no. 2, pp. 177–188, Jun. 2020.

[8] R. Alamanda, C. Suhery, and Y. Brianorman, "Aplikasi pendeteksi plagiat terhadap karya tulis berbasis web menggunakan natural language processing dan algoritma Knuth-Morris-Pratt," Jurnal Coding, Sistem Komputer Untan, vol. 4, no. 1, 2016.

[9] Y. Y. Pane and J. J. Sihombing, "Klasifikasi jenis burung menggunakan metode transfer learning," Jurnal Teknologi Terpadu, vol. 9, no. 2, pp. 89–94, 2023.

[10] R. Merdiansah and A. A. Ridha, "Analisis sentimen pengguna X Indonesia terkait kendaraan listrik menggunakan IndoBERT," Jurnal Ilmu Komputer dan Sistem Informasi (JIKOMSI), vol. 7, no. 1, pp. 221–228, 2024.

[11] H. F. Karim and A. P. Wibowo, "Kinerja metode fine-tuning IndoBERT untuk klasifikasi emosi multi-kelas pada teks informal bahasa Indonesia," Bulletin of Computer Science Research, vol. 6, no. 1, pp. 63–74, 2025.