JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9449
Hal 202-209
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Klasifikasi Komentar Toksik Berbahasa Indonesia di Media Sosial Berbasis Fine-Tuning IndoBERT

Luqman Nur Hakim*, Fida Maisa Hana, Widya Cholid Wahyudin

Fakultas Sains dan Teknologi, Program Studi Ilmu Komputer, Universitas Muhammadiyah Kudus, Kudus, Indonesia
Email: [1*]52023110014@std.umku.ac.id, [2]fidamaisa@umkudus.ac.id, [3,*] widyacholidwahyudin@umkudus.ac.id
Email Penulis Korespondensi: 52023110014@std.umku.ac.id
Submitted **31-12-2025;** Accepted **27-01-2026;** Published **28-02-2026**

Abstrak

Media sosial telah menjadi wadah utama bagi masyarakat Indonesia dalam berinteraksi dan bertukar informasi secara daring. Namun, kebebasan berekspresi di ruang digital sering kali disalahgunakan melalui penggunaan bahasa yang kasar, menghina, dan mengandung ujaran kebencian. Penelitian ini bertujuan untuk mengembangkan model klasifikasi komentar toksik berbahasa Indonesia dengan menggunakan arsitektur IndoBERT melalui proses fine-tuning. Model ini dipilih karena kemampuannya dalam memahami konteks semantik dua arah serta telah dilatih secara khusus menggunakan korpus Bahasa Indonesia, sehingga cocok untuk menangani gaya bahasa informal, singkatan, dan fenomena code-mixing yang umum di media sosial. Dataset yang digunakan adalah Indonesian Abusive and Hate Speech Twitter Text dengan total 12.942 data, terdiri atas 11.647 data latih dan 1.295 data validasi. Penelitian dilaksanakan secara daring menggunakan Google Colaboratory dengan dukungan GPU. Tahapan penelitian meliputi pra-pemrosesan data, tokenisasi, pelatihan model, dan evaluasi menggunakan metrik precision, recall, F1-score, serta confusion matrix. Hasil pengujian menunjukkan bahwa model IndoBERT mencapai performa tinggi dengan rata-rata precision sebesar 0,8842, recall sebesar 0,884, F1-score sebesar 0,883, dan akurasi sebesar 0,8834. Nilai tersebut menggambarkan keseimbangan performa antar kelas dan stabilitas model dalam mendeteksi komentar toksik maupun non-toksik. Penelitian ini berkontribusi terhadap pengembangan sistem moderasi konten otomatis berbahasa Indonesia, yang dapat diterapkan sebagai modul deteksi komentar berbasis API. Meskipun masih terbatas pada data Twitter dan klasifikasi biner, model ini memiliki potensi untuk dikembangkan ke arah klasifikasi multi-kelas dan lintas platform dalam mendukung ruang digital yang lebih aman dan sehat.

Kata Kunci: IndoBERT; Klasifikasi Teks; Ujaran Kebencian; Komentar Toksik; Pemrosesan Bahasa Alami

Abstract

Social media has become a primary platform for Indonesian society to interact and exchange information online. However, freedom of expression in digital spaces is often misused through the use of harsh, offensive, and hateful language. This study aims to develop a toxic comment classification model for the Indonesian language using the IndoBERT architecture through a fine-tuning process. IndoBERT was selected for its capability to understand bidirectional semantic context and its pretraining on a Bahasa Indonesia corpus, making it suitable for handling informal language styles, abbreviations, and common code-mixing phenomena in social media texts. The dataset used in this study is the Indonesian Abusive and Hate Speech Twitter Text, consisting of 12,942 entries 11,647 for training and 1,295 for validation. The research was conducted online using Google Colaboratory with GPU acceleration. The research stages included data preprocessing, tokenization, model training, and evaluation using precision, recall, F1-score, and confusion matrix as metrics. Evaluation results show that the fine-tuned IndoBERT model achieved high performance, with an average precision of 0.8842, recall of 0.884, F1-score of 0.883, and accuracy of 0.8834. These results indicate balanced performance across classes and strong model stability in detecting both toxic and non-toxic comments. This study contributes to the development of an automated Indonesian-language content moderation system, which can be deployed as a comment detection module via API. Although limited to Twitter data and binary classification, this model has the potential to be extended toward multi-class and cross-platform classification in supporting safer and healthier digital spaces in Indonesia.

Keywords: IndoBERT; Text Classification; Hate Speech; Toxic Comments; Natural Language Processing

1. PENDAHULUAN

Media sosial kini berperan sebagai ruang utama bagi masyarakat dalam melakukan komunikasi dan pertukaran informasi secara daring. Beragam platform berbasis internet memungkinkan pengguna mengekspresikan pendapat serta berinteraksi melalui media tulisan, gambar, maupun video secara terbuka. Di balik kemudahan tersebut, muncul berbagai permasalahan berupa penyebaran bahasa yang tidak pantas, termasuk penggunaan kata-kata kasar, penghinaan, serta komentar bermuatan kebencian atau toksik. Fenomena ini dapat berdampak pada kenyamanan pengguna sekaligus memengaruhi kualitas hubungan sosial dan suasana komunikasi di ruang digital Indonesia secara keseluruhan [1], [2], [3].

Berbagai laporan menunjukkan bahwa perilaku komunikasi negatif seperti penggunaan kata-kata kasar di media sosial semakin meningkat. Kementerian Komunikasi dan Informatika (Kominfo) melaporkan bahwa hingga September 2023 telah dilakukan penindakan terhadap lebih dari 3,7 juta konten bermuatan negatif di berbagai platform digital, termasuk ujaran kebencian dan pelecehan verbal [2]. [3] menyatakan bahwa komentar atau ujaran toksik di ruang digital memiliki dampak psikologis yang serius bagi korbannya serta berpotensi memicu konflik sosial di masyarakat. Apabila kondisi ini tidak dikendalikan, maka media sosial dapat kehilangan fungsinya sebagai sarana berbagi pengetahuan dan hiburan yang positif.

Upaya deteksi komentar toksik di media sosial Indonesia telah banyak dilakukan. Beberapa penelitian menggunakan metode pembelajaran mesin klasik seperti *Naïve Bayes* dan *Support Vector Machine* (SVM) untuk mengklasifikasikan unggahan maupun komentar pengguna berbahasa Indonesia [4], [5]. Meskipun metode tersebut

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9449
Hal 202-209
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

menunjukkan hasil yang cukup baik, berbagai tantangan linguistik seperti penggunaan bahasa daerah, ejaan tidak baku, singkatan, serta fenomena campur kode (*code-mixing*) masih menjadi hambatan dalam membangun sistem deteksi yang andal [6].

Dalam beberapa tahun terakhir, model berbasis transformer seperti BERT (*Bidirectional Encoder Representations from Transformers*) telah membawa kemajuan signifikan dalam bidang Pemrosesan Bahasa Alami (*Natural Language Processing* atau NLP) [7]. Untuk bahasa Indonesia, telah dikembangkan IndoBERT, yaitu model pralatih yang dilatih khusus pada korpus berbahasa Indonesia sehingga mampu memahami konteks kata secara lebih akurat [8]. Penelitian oleh [9] menunjukkan bahwa fine-tuning IndoBERT dapat meningkatkan performa klasifikasi pada data formal, namun belum secara mendalam mengeksplorasi konteks data informal dan toksik di media sosial [10]. Studi lain oleh [11] mengimplementasikan IndoBERT untuk analisis sentimen terhadap opini publik terkait kandidat presiden Indonesia. Namun, penelitian tersebut tidak secara khusus membahas deteksi ujaran kebencian. Hal ini menunjukkan bahwa pemanfaatan IndoBERT dalam konteks klasifikasi komentar toksik yang bercampur antara bahasa formal dan informal, termasuk slang dan bahasa alay, masih relatif minim.

Gap yang ingin dijawab oleh penelitian ini terletak pada keterbatasan eksplorasi model IndoBERT terhadap data komentar media sosial yang kompleks secara linguistik, serta belum adanya pendekatan yang secara sistematis menerapkan fine-tuning IndoBERT untuk klasifikasi komentar toksik berbahasa Indonesia dalam domain sosial digital yang dinamis. Penelitian ini memiliki kebaruan dalam mengevaluasi performa model pralatih tersebut dengan pendekatan yang disesuaikan terhadap karakteristik bahasa media sosial Indonesia, yang kaya akan variasi ekspresi dan nuansa lokal.

Berdasarkan permasalahan tersebut, penelitian ini memiliki kontribusi dalam menerapkan metode fine-tuning model IndoBERT untuk tugas klasifikasi komentar toksik berbahasa Indonesia di media sosial, dengan memanfaatkan kemampuan pemahaman konteks dua arah serta representasi semantik yang lebih baik. Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan sistem moderasi konten otomatis yang lebih efektif di Indonesia.

Adapun tujuan dari penelitian ini adalah untuk menganalisis dan mengevaluasi kinerja model IndoBERT yang telah melalui proses fine-tuning dalam mendeteksi komentar toksik pada media sosial berbahasa Indonesia. Penelitian ini juga membuka peluang pengembangan lebih lanjut berupa penerapan model fine-tuning IndoBERT dalam sistem moderasi komentar otomatis, misalnya pada platform Discord. Dengan pendekatan ini, model dapat diimplementasikan untuk mendeteksi komentar toksik secara *real-time* dan mendukung terciptanya ekosistem komunikasi digital yang lebih sehat.

2. METODOLOGI PENELITIAN

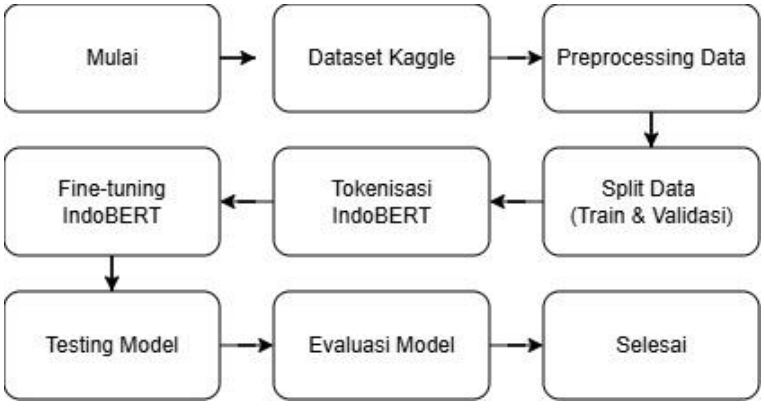
2.1 Desain Penelitian dan Pendekatan

Penelitian ini menggunakan pendekatan kuantitatif dengan jenis eksperimen, yang bertujuan untuk menguji efektivitas fine-tuning IndoBERT dalam klasifikasi komentar toksik pada media sosial berbahasa Indonesia. Pendekatan ini dipilih karena sesuai dengan karakteristik penelitian machine learning yang mengandalkan proses pelatihan model terhadap data berlabel, serta mengukur performa berdasarkan metrik kuantitatif seperti akurasi, presisi, dan recall.

Model IndoBERT adalah adaptasi dari arsitektur BERT (*Bidirectional Encoder Representations from Transformers*) yang telah dilatih secara khusus menggunakan korpus bahasa Indonesia [8]. Dengan menggunakan pendekatan fine-tuning, model dapat dioptimalkan ulang agar lebih sensitif terhadap karakteristik linguistik unik dari teks-teks media sosial yang kerap mengandung bahasa tidak baku, singkatan, dan fenomena *code-mixing* [12].

2.2 Tahapan Penelitian dan Alur Fine-Tuning IndoBERT

Untuk menjamin keberhasilan pengembangan model, tahapan penelitian disusun berdasarkan alur logis yang terdiri dari akuisisi data, pra-pemrosesan, tokenisasi, pelatihan model, evaluasi, dan analisis kinerja. Alur proses tersebut divisualisasikan dalam Gambar 1 berikut, yang memberikan gambaran menyeluruh mengenai langkah-langkah penelitian yang dilakukan.



Gambar 1. Diagram Alur Proses Fine-Tuning IndoBERT untuk Klasifikasi Komentar Toksik

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9449
Hal 202-209
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

2.2.1 Akuisisi Data: Dataset Kaggle

Dataset yang digunakan adalah dataset publik bertajuk Indonesian *Abusive and Hate Speech Twitter Text*, tersedia di platform Kaggle [13]. Dataset ini terdiri dari ribuan entri komentar yang dikategorikan ke dalam beberapa label seperti *abusive, hate speech*, dan normal. Untuk kebutuhan penelitian ini, enam label asli dikonversi menjadi dua kelas utama: toxic (1) dan non-toxic (0).
Penggunaan dataset ini dianggap tepat karena:
a. Relevan dengan konteks komentar sosial media di Indonesia.
b. Telah melalui proses kurasi awal oleh penyedia dataset.
c. Bersifat terbuka (open source), memudahkan replikasi dan validasi.

2.2.2 Pra-pemrosesan Data

Pra-pemrosesan menjadi tahap penting dalam pemrosesan bahasa alami (NLP), karena kualitas input akan sangat memengaruhi performa model. Adapun tahapan pra-pemrosesan meliputi:
a. Penghapusan entri kosong dan duplikat, untuk menghindari bias.
b. Normalisasi teks, seperti konversi ke huruf kecil, penghapusan tanda baca, emotikon, mention (@username), hashtag, dan URL.
c. Label binarisasi, yaitu konversi enam label ke dalam dua kelas utama sesuai kebutuhan klasifikasi biner.
Hasil akhir tahap ini adalah data bersih yang siap dibagi menjadi data latih dan data validasi.

2.2.3 Pembagian Data: Split Train & Validasi

Dataset yang telah diproses kemudian dibagi menjadi dua subset: 90% untuk pelatihan (*training*) dan 10% untuk validasi (*validation*) menggunakan metode *stratified split*. Teknik ini memastikan bahwa proporsi antara kelas *toxic* dan *non-toxic* tetap seimbang pada kedua subset [14].

Stratifikasi penting dilakukan karena distribusi data sosial media umumnya tidak seimbang komentar toksik lebih sedikit dibanding yang normal dan ketidakseimbangan ini dapat menyebabkan bias model dalam klasifikasi.

2.2.4 Tokenisasi IndoBERT

Tokenisasi dilakukan menggunakan *tokenizer* milik IndoBERT-base-p1, yang menghasilkan dua tensor utama: input_ids dan attention_mask. Tokenisasi dilakukan dengan parameter berikut:
a. max_length: 128
b. padding: max_length
c. truncation: true

Tokenisasi ini memastikan seluruh teks memiliki panjang input yang seragam dan optimal untuk diproses dalam arsitektur transformer [15].

2.2.5 Fine-tuning Model IndoBERT

Fine-tuning merupakan proses pelatihan lanjutan terhadap model pralatih dengan data spesifik. Model IndoBERT yang digunakan adalah AutoModelForSequenceClassification dari HuggingFace Transformers, dengan konfigurasi sebagai berikut:
a. Epochs: 3
b. Batch size: 16
c. Optimizer: AdamW
d. Learning rate: 5e-5
e. Evaluation step: setiap 500 langkah
f. Checkpoint: disimpan berdasarkan nilai F1-score tertinggi

Fine-tuning dilakukan dengan menggunakan Google Colaboratory yang menyediakan akselerasi GPU untuk efisiensi waktu dan kinerja [16].

2.2.6 Pengujian dan Evaluasi Model

Setelah proses pelatihan (*fine-tuning*) selesai, model dievaluasi menggunakan subset data validasi yang telah dipisahkan dari data awal dengan rasio 90:10. Evaluasi ini bertujuan untuk mengukur efektivitas model IndoBERT dalam mengklasifikasikan komentar toksik berbahasa Indonesia secara akurat.
Adapun metrik evaluasi yang digunakan meliputi:
a. Akurasi (*accuracy*): Mengukur proporsi klasifikasi yang benar terhadap seluruh data.
b. Presisi (*precision*): Mengukur ketepatan model dalam mengidentifikasi komentar toxic tanpa memasukkan yang non-toxic.
c. Recall: Mengukur sejauh mana model berhasil mendeteksi seluruh komentar toxic yang sebenarnya ada.
d. F1-score: Rata-rata harmonik antara precision dan recall, memberikan penilaian menyeluruh terhadap performa klasifikasi.

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9449
Hal 202-209
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

e. Confusion matrix: Evaluasi akhir juga dilengkapi dengan confusion matrix, yaitu matriks 2x2 yang merepresentasikan distribusi prediksi benar dan salah untuk dua kelas target, yakni toxic dan non-toxic. Tabel 1 berikut menyajikan hasil evaluasi model IndoBERT berdasarkan data validasi.

Tabel 1. Confusion Matrix Hasil Evaluasi Model IndoBERT

	Prediksi: Non-Toxic (0)	Prediksi: Toxic (1)
Kondisi Aktual: Non-Toxic (0)	514 (Benar Negatif)	64 (Salah Positif)
Kondisi Aktual: Toxic (1)	59 (Salah Negatif)	666 (Benar Positif)

Berdasarkan Tabel 1, model berhasil mengklasifikasikan 514 komentar non-toxic secara benar (*true negative*) dan 666 komentar *toxic* secara benar (*true positive*). Namun, masih terdapat 64 komentar *non-toxic* yang salah diklasifikasikan sebagai toxic (*false positive*), serta 59 komentar *toxic* yang tidak terdeteksi (*false negative*). Distribusi ini menunjukkan bahwa model memiliki performa klasifikasi yang cukup seimbang, dengan kecenderungan menghasilkan nilai recall dan precision yang tinggi.

Evaluasi dilakukan setiap 500 langkah pelatihan (*evaluation step*), dan model terbaik ditentukan berdasarkan nilai F1-score tertinggi selama proses pelatihan. Pemilihan metrik-metrik tersebut merujuk pada praktik terbaik dalam evaluasi tugas klasifikasi teks pendek di bidang pemrosesan bahasa alami (*Natural Language Processing*) [17].

2.3 Etika Penelitian

Penelitian ini tidak melibatkan subjek manusia secara langsung. Dataset yang digunakan bersifat publik dan telah dianonimkan oleh penyedia data. Peneliti memastikan seluruh proses mematuhi prinsip etika penelitian dan perlindungan data pribadi digital. Hasil penelitian diharapkan dapat berkontribusi dalam pengembangan sistem deteksi komentar toksik yang adil, transparan, dan tidak diskriminatif [18].

3. HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil dan pembahasan dari penelitian pengembangan model klasifikasi komentar toksik berbahasa Indonesia menggunakan IndoBERT. Analisis mencakup tahapan pra-pemrosesan data, tokenisasi, fine-tuning model, pengujian performa, serta evaluasi hasil berdasarkan metrik dan perbandingan dengan penelitian terdahulu.

3.1 Dataset dan Tahap Tokenisasi

Dataset yang digunakan dalam penelitian ini adalah *Indonesian Abusive and Hate Speech Twitter Text* [13], yang berisi unggahan pengguna Twitter berbahasa Indonesia dengan berbagai kategori ujaran toksik, seperti HS (*Hate Speech*), *Abusive*, dan *HS_Group*.

Dalam penelitian ini, seluruh label multi-kelas tersebut digabungkan menjadi satu label biner, yaitu *non-toxic* (0) dan *toxic* (1), untuk mendukung pendekatan klasifikasi biner menggunakan model IndoBERT. Dataset ini dipilih karena mencerminkan karakteristik bahasa informal, penggunaan singkatan, serta fenomena *code-mixing* yang umum di media sosial.

Tabel 2. Dataset "Indonesian Abusive and Hate Speech Twitter Text"

Tweet	HS	Abusive	HS_Group	Toxic Label
disaat semua cowok berusaha melacak perhatian gue...	1	1	0	1
RT USER: siapa yang telat ngasih tau elu? edan sarap...	0	1	0	1
Kadang aku berfikir, kenapa aku tetap percaya...	0	0	0	0
USER USER Kaum cebong kapir udah keliatan dongoknya...	1	1	1	1

Setiap komentar melalui tahap preprocessing dengan menghapus karakter non-alfabet, tanda baca, dan mention pengguna, serta melakukan normalisasi huruf kecil. Kolom Toxic Label digunakan sebagai label akhir dalam proses pelatihan model klasifikasi.

Sebelum pelatihan dilakukan, setiap teks dikonversi menjadi token menggunakan IndoBERT tokenizer. Tokenisasi ini memecah teks menjadi satuan kata dan sub-kata agar dapat diproses oleh model transformer.

Tabel 3. Visualisasi Tokenisasi Menggunakan IndoBERT

Index	Teks	Tokens	Length
1	USER USER USER si mahfud sekarang beda, sptnya uda pasang badan utk rezim ini, kalau dulu terkenal sangat kritis.'	[user, user, user, si, mahfud, sekarang, beda, ,, spt, ##nya, uda, pasang, badan, utk, rezim, ini, ,, kalau, dulu, terkenal, sangat, kritis, ., ']	24
2	Agak horror juga dapet smirk dari orang asing'	[agak, hor, ##ror, juga, dapet, sm, ##ir, ##k, dari, orang, asing, ']	37

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9449
Hal 202-209
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

3	USER klo seiman nipu hukumnya halal ga mgkin didemo kaum Bani micin'	[user, klo, sei, ##man, nip, ##u, hukumnya, halal, ga, mg, ##kin, did, ##emo, kaum, bani, mic, ##in,']	18

Tokenisasi menghasilkan dua keluaran utama, yaitu input_ids dan attention_mask, yang kemudian dikonversi menjadi tensor sebelum digunakan dalam tahap fine-tuning IndoBERT.

3.2 Implementasi dan Evaluasi Model

3.2.1 Tahapan Implementasi

Dataset hasil pra-pemrosesan (lihat Subbab 3.1) digunakan sebagai input untuk proses *fine-tuning* model IndoBERT. Model ini merupakan arsitektur berbasis *transformer* yang telah dilatih secara khusus pada korpus berbahasa Indonesia [19].

Sebelum proses pelatihan dimulai, dataset dibagi menjadi dua bagian, yaitu 11.647 data latih dan 1.295 data validasi, menggunakan teknik *stratified split* dengan rasio 90:10. Teknik ini dipilih untuk menjaga proporsi yang seimbang antara kelas *toxic* dan non-*toxic* pada kedua subset data, sehingga performa model tetap adil dalam mengenali kedua kelas [14].

Tokenisasi dilakukan menggunakan tokenizer IndoBERT dengan konfigurasi max_length=128, padding=max_length, dan truncation=True. Data hasil tokenisasi kemudian dikonversi menjadi tensor agar dapat diproses oleh arsitektur transformer.

Model yang digunakan adalah IndoBERT versi base yang di-fine-tune menggunakan arsitektur AutoModelForSequenceClassification dari pustaka Transformers [15]. Seluruh proses pelatihan dilakukan di lingkungan Google Colaboratory dengan akselerasi GPU, guna meningkatkan efisiensi waktu komputasi [16].
Konfigurasi parameter pelatihan model disajikan dalam Tabel 4 berikut.

Tabel 4. Konfigurasi Parameter Pelatihan Model IndoBERT

Parameter	Nilai
Jumlah epoch	3
Batch size	16
Optimizer	AdamW
Learning rate	5e-5
Evaluasi	setiap 500 langkah
Checkpoint terbaik	berdasarkan nilai F1-score tertinggi

3.2.2 Hasil Evaluasi Kinerja Model

Selama proses pelatihan, nilai loss menurun secara bertahap, sementara F1-score meningkat pada setiap epoch. Hal ini menunjukkan bahwa model belajar secara progresif terhadap pola linguistik komentar berbahasa Indonesia [20], [21].
Evaluasi model dilakukan menggunakan empat metrik utama: Precision, Recall, F1-score, dan Accuracy.

Tabel 5. Hasil Evaluasi Kinerja Model IndoBERT

Kelas	Precision	Recall	F1-score	Jumlah Data
0 (Non-toxic)	0,9127	0,857	0,884	671
1 (Toxic)	0,8556	0,912	0,883	624
Rata-rata	0,8842	0,884	0,883	1.295

Tabel 5 menunjukkan hasil evaluasi kinerja model IndoBERT terhadap data validasi. Kelas non-toxic memiliki nilai precision sebesar 0,9127 dan recall sebesar 0,857, menunjukkan bahwa model cukup akurat dalam mengidentifikasi komentar yang tidak bersifat toksik, meskipun masih terdapat sejumlah false positive. Sebaliknya, kelas toxic memperoleh precision sebesar 0,8556 dan recall sebesar 0,912, yang menunjukkan kemampuan model yang sangat baik dalam menangkap komentar toksik meskipun dengan kemungkinan adanya false negative.

Nilai F1-score yang seimbang di kedua kelas (±0,883) mengindikasikan bahwa model tidak bias terhadap salah satu kelas. Hasil ini mendukung temuan dari [10], yang menekankan pentingnya keseimbangan antara precision dan recall dalam tugas deteksi ujaran kebencian menggunakan model berbasis transformer.

Rata-rata keseluruhan metrik juga menunjukkan bahwa model mampu beradaptasi dengan baik terhadap kompleksitas bahasa informal dan fenomena code-mixing yang sering dijumpai dalam komunikasi di media sosial [12].

Tabel 6. Confusion Matrix Model IndoBERT

Kelas Aktual / Prediksi	Non-toxic	Toxic
Non-toxic	575	96
Toxic	55	569

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 1, Februari 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i1.9449
Hal 202-209
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Tabel 6 menyajikan confusion matrix sebagai representasi distribusi hasil klasifikasi model. Dari total 1.295 data validasi, model berhasil mengklasifikasikan 575 komentar non-toxic dan 569 komentar toxic secara benar. Namun demikian, terdapat 96 komentar non-toxic yang diklasifikasikan secara salah sebagai toxic (false positive), dan 55 komentar toxic yang tidak terdeteksi (false negative).

Tingginya jumlah true positive dan true negative mencerminkan kemampuan generalisasi model yang baik terhadap kedua kelas. Akan tetapi, nilai false positive yang relatif tinggi pada kelas non-toxic menunjukkan kecenderungan model untuk bersikap konservatif, yaitu lebih memilih mengklasifikasikan komentar sebagai toxic demi meminimalkan risiko konten berbahaya lolos dari deteksi. Strategi seperti ini cukup umum dalam sistem moderasi berbasis NLP, seperti dijelaskan oleh [12].

Tabel 7. Statistik Deskriptif F1-score Antar Kelas

Parameter	Nilai
Mean	0,8834
Standar Deviasi	0,0007
Median	0,8834
Minimum	0,8829
Maksimum	0,8839

Tabel 7 menyajikan statistik deskriptif dari F1-score antar kelas. Nilai rata-rata (mean) sebesar 0,8834 dan median yang identik menunjukkan distribusi kinerja model yang simetris. Rentang nilai minimum (0,8829) dan maksimum (0,8839) yang sangat kecil, serta nilai standar deviasi hanya 0,0007, mengindikasikan bahwa model memiliki stabilitas kinerja yang sangat tinggi antar kelas.

Variasi F1-score yang minimal ini memperkuat hasil sebelumnya bahwa model mampu melakukan klasifikasi dengan tingkat akurasi yang konsisten, tanpa adanya gejala overfitting pada salah satu kelas. Hasil ini sejalan dengan temuan [8], yang menunjukkan bahwa model IndoBERT memiliki kestabilan tinggi dalam klasifikasi teks pendek setelah dilakukan fine-tuning secara tepat.

3.3 Analisis Dan Pembahasan

Berdasarkan hasil pengujian yang telah dilakukan, fine-tuning IndoBERT terbukti mampu meningkatkan kemampuan model dalam mengenali ujaran toksik dan bahasa kasar berbahasa Indonesia. Nilai F1-score rata-rata sebesar 0,883 menunjukkan keseimbangan yang baik antara precision dan recall.

Keunggulan ini berasal dari mekanisme perhatian dua arah (*bidirectional attention*) pada arsitektur IndoBERT, yang memungkinkan model memahami makna kata berdasarkan konteks kalimat secara menyeluruh [7], [22].

Tahapan pra-pemrosesan data yang bersih turut berkontribusi signifikan terhadap hasil, karena mengurangi noise linguistik dan meningkatkan kemampuan model dalam melakukan generalisasi [14], [23].

Namun demikian, model masih menghadapi kesulitan dalam mendeteksi ujaran bersifat sarkastik atau implisit, terutama pada teks yang mengandung *code-mixing*. Fenomena ini juga ditemukan oleh [24], yang menyoroti pentingnya penggabungan *contextual embeddings* dengan *emotion detection* untuk menangani kasus seperti ini.

Kebaruan penelitian ini terletak pada penerapan fine-tuning IndoBERT untuk komentar media sosial berbahasa Indonesia dengan gaya informal dan campur kode, berbeda dengan sebagian besar penelitian terdahulu yang fokus pada teks formal seperti berita atau artikel opini [12], [25]

Dari sisi penerapan, hasil penelitian ini membuka peluang implementasi model dalam sistem moderasi komentar otomatis, misalnya di platform seperti Discord. Dengan integrasi melalui API, model dapat mendeteksi komentar toksik secara real-time dan membantu moderator manusia dalam memelihara lingkungan komunikasi digital yang aman [26], [27].

Konfigurasi parameter sederhana yang digunakan (3 epoch, batch size 16, learning rate 5e-5) terbukti memberikan hasil optimal tanpa memerlukan sumber daya GPU yang besar. Hasil ini sejalan dengan temuan [12], yang menunjukkan efisiensi IndoBERT-base pada tugas klasifikasi teks berbahasa Indonesia.

Walaupun hasilnya tinggi, penelitian ini memiliki keterbatasan, yaitu dataset yang digunakan hanya berasal dari Twitter. Variasi gaya bahasa di platform lain seperti YouTube, TikTok, atau Instagram belum tercakup. Penelitian selanjutnya disarankan untuk menerapkan augmentasi data berbasis GPT [10] dan multi-task learning, agar model mampu mengenali tingkat toksisitas secara lebih detail (*fine-grained classification*).

Selain itu, penerapan teknik attention visualization seperti yang dikembangkan [25] juga direkomendasikan agar model lebih transparan dan mendukung prinsip Explainable AI (XAI) dalam sistem moderasi konten digital.

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa IndoBERT dengan *fine-tuning* mampu memberikan kinerja tinggi, sebagaimana ditunjukkan oleh nilai F1-score sebesar 0,883 (lihat Tabel 5), yang menggambarkan keseimbangan antara precision dan recall pada kedua kelas. Nilai ini berada dalam kategori performa sangat baik untuk tugas klasifikasi teks pendek dalam domain media sosial, sebagaimana dijelaskan oleh [17], bahwa F1-score di atas 0,80

umumnya mencerminkan kemampuan model yang optimal dalam mengatasi ketidakseimbangan data dan variasi linguistik yang kompleks. Selain itu, model menunjukkan stabilitas performa antar kelas dan efisiensi komputasi yang baik, menjadikannya layak untuk diintegrasikan dalam sistem moderasi konten otomatis guna menciptakan ruang digital yang lebih aman, sehat, dan produktif di Indonesia.

Dengan tahapan implementasi yang sistematis mulai dari preprocessing hingga evaluasi, serta konfigurasi pelatihan yang efisien, model IndoBERT berhasil menunjukkan performa tinggi dan stabil. Penelitian ini tidak hanya mengonfirmasi efektivitas arsitektur transformer dalam tugas NLP berbahasa Indonesia, tetapi juga memperlihatkan potensi praktisnya dalam pengembangan sistem moderasi konten berbasis AI untuk platform digital yang dinamis.

4. KESIMPULAN

Penelitian ini berhasil mengembangkan model klasifikasi komentar toksik berbahasa Indonesia dengan memanfaatkan arsitektur IndoBERT yang telah melalui proses fine-tuning terhadap dataset publik Indonesian Abusive and Hate Speech Twitter Text. Model ini dirancang untuk mendeteksi dan mengklasifikasikan komentar media sosial ke dalam dua kelas utama, yaitu toksik (1) dan non-toksik (0). Hasil pengujian menunjukkan bahwa model IndoBERT yang dikembangkan mampu mencapai performa yang sangat baik dengan rata-rata precision sebesar 0,8842, recall sebesar 0,884, dan F1-score sebesar 0,883. Nilai tersebut mencerminkan keseimbangan performa antara kedua kelas serta menunjukkan bahwa model tidak mengalami bias klasifikasi yang signifikan. Analisis confusion matrix memperlihatkan bahwa sebagian besar komentar berhasil diklasifikasikan dengan benar, sedangkan analisis statistik deskriptif memperkuat temuan stabilitas model dengan selisih F1-score antar kelas yang sangat kecil. Keunggulan performa model ini dipengaruhi oleh kemampuan arsitektur transformer dalam memahami konteks semantik dua arah (*bidirectional semantic representation*), serta efektivitas tahap pra-pemrosesan yang mampu meminimalkan noise linguistik pada data pelatihan. Secara keseluruhan, penelitian ini memberikan kontribusi nyata dalam pengembangan sistem moderasi konten otomatis berbahasa Indonesia, yang dapat dimanfaatkan sebagai sistem penyaringan awal (*pre-filtering system*) untuk mendeteksi komentar berpotensi berbahaya di platform digital. Meskipun demikian, penelitian ini masih memiliki beberapa keterbatasan, khususnya pada penggunaan dataset tunggal yang hanya bersumber dari platform Twitter dan pendekatan klasifikasi yang masih bersifat biner. Untuk pengembangan selanjutnya, disarankan agar penelitian mendatang memperluas cakupan dataset lintas domain, menerapkan pendekatan *multi-class classification*, serta mengintegrasikan teknik *cross-domain transfer learning* atau real-time moderation system, sehingga model mampu beradaptasi secara dinamis terhadap keragaman konteks dan evolusi bahasa Indonesia di ranah digital. Selain itu, hasil penelitian ini dapat dikembangkan lebih lanjut sebagai modul deteksi komentar toksik berbasis API yang diintegrasikan dengan sistem moderasi teks pada platform interaktif seperti Discord atau YouTube Live Chat, guna meningkatkan keamanan komunikasi digital secara *real-time* dan mendukung ekosistem media sosial yang lebih sehat di Indonesia.

REFERENCES

[1] S. Kemp, “Digital 2024: Indonesia,” DataReportal. Accessed: Dec. 06, 2025. [Online]. Available: https://datareportal.com/reports/digital-2024-indonesia

[2] Kominfo, “Kominfo tangani 3,7 juta konten negatif hingga 17 September 2023,” Kontan.co.id. Accessed: Dec. 06, 2025. [Online]. Available: https://nasional.kontan.co.id/news/kominfo-tangani-37-juta-konten-negatif-hingga-17-september-2023

[3] H. Rahmi and A. Corsini, “Tinjauan Fenomena ‘Hate Speech’ dengan Muatan Politik di Indonesia dalam Perspektif ‘ Psychological Hatred’ Review of the phenomenon of ‘Hate Speech’ with political content in Indonesia in ‘Psychological Hatred’ perspective,” 2020.

[4] D. P. N. Lyrawati, “Deteksi Ujaran Kebencian pada Twitter Menjelang Pilpres 2019 dengan Machine Learning,” MATHunesa: Jurnal Ilmiah Matematika, vol. 7, no. 3, pp. 206–211, 2019.

[5] R. M. Yazid, F. R. Umbara, and P. N. Sabrina, “Deteksi Ujaran Kebencian dengan Metode Klasifikasi Naive Bayes dan Metode N-Gram pada Dataset Multi-Label Twitter Berbahasa Indonesia,” Informatics and Digital Expert (INDEX), vol. 4, no. 2, pp. 46–52, Nov. 2022, [Online]. Available: http://index.unper.ac.id

[6] I. Budi, Analisis Media Sosial Sebagai Upaya Dini Deteksi Potensi Konflik Masyarakat di Dunia Maya. Depok: Fakultas Ilmu Komputer, Universitas Indonesia, 2023.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), Association for Computational Linguistics (ACL), May 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[8] B. Wilie et al., “IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 843–857. Accessed: Dec. 06, 2025. [Online]. Available: https://aclanthology.org/2020.aacl-main.85/

[9] G. Z. Nabiilah, S. Y. Prasetyo, Z. N. Izdihar, and A. S. Girsang, “BERT base model for toxic comment analysis on Indonesian social media,” Procedia Comput. Sci., vol. 216, pp. 714–721, Jan. 2023, doi: 10.1016/J.PROCS.2022.12.188.

[10] E. W. Pamungkas, D. G. P. Putri, and A. Fatmawati, "Hate Speech Detection in Bahasa Indonesia: Challenges and Opportunities," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 6, pp. 1175–1181, May 2023, Accessed: Dec. 06, 2025. [Online]. Available: https://doi.org/10.14569/IJACSA.2023.01406125

[11] P. Sayarizki, Hasmawati, and H. Nurrahmi, "Implementation of IndoBERT for Sentiment Analysis of Indonesian Presidential Candidates," Indonesian Journal on Computing (Indo-JC), vol. 9, no. 2, pp. 61–72, Aug. 2024, doi: 10.34818/INDOJC.2024.9.2.934.

[12] M. O. Ibrohim and I. Budi, "Hate speech and abusive language detection in Indonesian social media: Progress and challenges," Heliyon, vol. 9, no. 8, p. e18647, Aug. 2023, doi: 10.1016/J.HELIYON.2023.E18647.

[13] I. F. Putra, "Indonesian Abusive and Hate Speech Twitter Text," Kaggle. Accessed: Dec. 06, 2025. [Online]. Available: https://www.kaggle.com/datasets/ilhamfp31/indonesian-abusive-and-hate-speech-twitter-text

[14] K. Kamdan, M. P. Anugrah, M. J. Almutaali, R. Ramdani, and I. L. Kharisma, "Performance Analysis of IndoBERT for Detection of Online Gambling Promotion in YouTube Comments," in 7th International Global Conference Series on ICT Integration in Technical Education & Smart Society, Aizuwakamatsu, Japan: MDPI AG, Sep. 2025, p. 66. doi: 10.3390/engproc2025107066.

[15] T. Wolf et al., "HuggingFace's Transformers: State-of-the-art Natural Language Processing," Journal of Machine Learning Research, Jul. 2020, [Online]. Available: http://arxiv.org/abs/1910.03771

[16] Google, "Google Colaboratory," Google. Accessed: Dec. 06, 2025. [Online]. Available: https://colab.research.google.com/

[17] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," Sci. Rep., vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-56706-x.

[18] I. P. G. H. Suputra, Linawati, G. Sukadarmika, N. P. Sastra, N. M. A. Wilani, and I. M. A. Setiawan, "Improved Cognitive Distortion Detection using IndoBERT and Important Words Approach for Bahasa Indonesia," International Journal on Informatics Visualization, vol. 9, no. 6, pp. 2272–2278, 2025, doi: 10.62527/joiv.9.6.3576.

[19] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020), International Committee on Computational Linguistics, 2020, pp. 757–770. Accessed: Dec. 06, 2025. [Online]. Available: https://aclanthology.org/2020.coling-main.66

[20] A. F. Hidayatullah, R. A. Apong, D. T. C. Lai, and A. Qazi, "Pre-trained language model for code-mixed text in Indonesian, Javanese, and English using transformer," Soc. Netw. Anal. Min., vol. 15, no. 1, Dec. 2025, doi: 10.1007/s13278-025-01444-9.

[21] P. A. Mufva, K. H. Chandra, K. F. Aji, I. A. Iswanto, and S. Joddy, "Performance comparison of deep learning approaches for Indonesian twitter hate speech detection using IndoBERTweet embedding," Procedia Comput. Sci., vol. 269, pp. 1663–1671, Jan. 2025, doi: 10.1016/J.PROCS.2025.09.109.

[22] A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS 2017), Curran Associates, Inc., 2017, pp. 5998–6008. doi: 10.5555/3295222.3295349.

[23] Y. Findawati, D. Purwitasari, and A. B. Raharjo, "Dangerous Speech Classification on Twitter Using Multilabel Aspects and Structural Features," IEEE Access, vol. 13, pp. 189506–189527, 2025, doi: 10.1109/ACCESS.2025.3627932.

[24] I. Amal and E. W. Pamungkas, "Enhancing Hate Speech Detection in Indonesia Code-Mixed Tweets: the Role of Oversampling and Undersampling Techniques," in 2025 International Conference on Smart Computing, IoT and Machine Learning (SIML), 2025, pp. 1–6. doi: 10.1109/SIML65326.2025.11081166.

[25] A. Fitro, A. Praba Ristadi Pinem, O. Saeful Bachri, and Chartini, "Cyberbullying Detection on Instagram using IndoBERTa Model," International Conference on Digital Business Innovation and Technology Management (ICONBIT), vol. 1, no. 2, Aug. 2025, [Online]. Available: https://proceeding.unesa.ac.id/index.php/iconbit/article/view/5798

[26] R. A. Saputra and Y. Sibaroni, "Multilabel Hate Speech Classification in Indonesian Political Discourse on X using Combined Deep Learning Models with Considering Sentence Length," Jurnal Ilmu Komputer dan Informasi, vol. 18, no. 1, pp. 113–125, Feb. 2025, doi: 10.21609/jiki.v18i1.1440.

[27] V. K. Santoso, N. C. Purba, C. A. Herli, and Y. Muliono, "Privacy Classification of Indonesian Chat Logs Using Indobert Model Variants," in 2025 International Seminar on Intelligent Technology and Its Applications (ISITIA), 2025, pp. 136–141. doi: 10.1109/ISITIA66279.2025.11137460.