

Perbandingan Embedding IndoBERT dan FastText dalam Deteksi Cyberbullying Berbahasa Indonesia

Comparison of IndoBERT and FastText Embeddings in Indonesian Cyberbullying Detection

Muhammad Ihksan Alfiansyah, Diana*

Fakultas Sains Teknologi, Program Studi Teknik Informatika, Universitas Bina Darma, Palembang, Indonesia

Email: ¹muhammadihksanalfiansyah@gmail.com, ^{2,*}diana@binadarma.ac.id

Email Penulis Korespondensi: diana@binadarma.ac.id

Received: 27/08/2026, Revised: 16/09/2026, Accepted: 21/09/2026, Published: 22/09/2026

Abstrak—Eskalasi cyberbullying di media sosial menuntut adanya sistem moderasi konten otomatis yang mampu mengurai karakteristik bahasa digital tidak baku. Penelitian ini membandingkan efektivitas contextual word embedding (IndoBERT) dan static subword embedding (FastText) dalam klasifikasi teks multi-kelas (Aman, Ringan, Sedang, Tinggi, dan Sangat Berbahaya) pada komentar media sosial berbahasa Indonesia. Batasan penelitian difokuskan pada analisis arsitektur embedding, di mana IndoBERT dipadukan dengan Dense Layer, sedangkan FastText dievaluasi menggunakan Logistic Regression dan Random Forest (15 trees). Korpus data berjumlah 16.643 sampel diekstraksi dari platform media sosial dan diproses melalui tahapan pembersihan teks serta normalisasi slang berbasis kamus leksikal. Hasil pengujian menunjukkan bahwa arsitektur kontekstual dua arah IndoBERT memberikan kinerja tertinggi dengan Akurasi 82,15% dan Macro F1-Score 74,97%. Pada representasi FastText, pendekatan linear Logistic Regression (Akurasi 78,70%) tampil lebih stabil dibandingkan model Random Forest (Akurasi 65,50%). Hasil ini menunjukkan bahwa contextual embedding lebih adaptif dalam membedakan sarkasme dan leksikon bermutasi pada teks digital dibandingkan static embedding. Sebagai kontribusi penelitian, studi ini tidak hanya menawarkan arsitektur klasifikasi lima tingkat keparahan yang presisi untuk bahasa Indonesia informal, tetapi juga menyediakan purwarupa sistem moderasi otomatis (CyberGuard AI) yang siap diimplementasikan untuk mitigasi risiko secara waktu-nyata.

Kata Kunci: Cyberbullying; IndoBERT; FastText; Multi-class Classification; Natural Language Processing; Transformer

Abstract—The escalation of cyberbullying on social media necessitates automated content moderation systems capable of deciphering the characteristics of non-standard digital language. This study compares the effectiveness of contextual word embeddings (IndoBERT) and static subword embeddings (FastText) in multi-class text classification (Safe, Mild, Moderate, Severe, and Very Dangerous) for Indonesian-language social media comments. The research focuses on analyzing embedding architectures: IndoBERT is integrated with a Dense Layer, while FastText is evaluated using Logistic Regression and Random Forest (15 trees). A dataset of 16,643 samples was extracted from social media platforms and processed through text cleaning and dictionary-based slang normalization stages. Test results indicate that the bidirectional contextual architecture of IndoBERT achieved the highest performance, with an accuracy of 82.15% and a Macro F1-Score of 74.97%. Regarding FastText representations, the linear Logistic Regression approach (78.70% accuracy) proved more stable than the Random Forest model (65.50% accuracy). These findings demonstrate that contextual embeddings are more adaptive than static embeddings at distinguishing sarcasm and mutated lexicons in digital text. As a research contribution, this study not only offers a precise five-level severity classification architecture for informal Indonesian but also provides a prototype automated moderation system (CyberGuard AI) ready for implementation in real-time risk mitigation.

Keywords: Cyberbullying; IndoBERT; FastText; Multi-class Classification; Natural Language Processing; Transformer

1. PENDAHULUAN

Pesatnya perkembangan teknologi informasi dan tingginya penetrasi internet di Indonesia telah mentransformasikan media sosial menjadi ruang publik digital utama dalam pola komunikasi masyarakat. Platform daring seperti Twitter/X dan Instagram memang memfasilitasi pertukaran gagasan secara terbuka, interaktif, dan waktu-nyata (*real-time*). Namun di sisi lain, kebebasan berinteraksi dan minimnya batasan di ruang virtual ini membawa dampak negatif, salah satunya adalah maraknya kemunculan perilaku perundungan siber (*cyberbullying*) yang sangat rentan terjadi di kalangan remaja [1]. Platform daring seperti Twitter/X dan Instagram memfasilitasi pertukaran gagasan secara terbuka, interaktif, dan waktu-nyata (*real-time*). Kendati demikian, kebebasan berekspresi di ruang siber kerap disalahgunakan hingga memicu eskalasi fenomena *cyberbullying* (perundungan siber) di kolom komentar, seperti penghinaan personal, pelecehan verbal, fitnah, dan ujaran kebencian berbasis identitas kelompok. Paparan toksisitas digital ini berdampak destruktif terhadap stabilitas psikologis korban, mulai dari stres kronis, kecemasan akut, degradasi harga diri (*self-esteem*), depresi berat, penarikan diri dari lingkungan sosial, hingga risiko fatal berupa ideasi bunuh diri [2]. Fenomena agresif ini diperkuat oleh *online disinhibition effect*, khususnya *toxic disinhibition*, di mana kelonggaran kontrol sosial, anonimitas akun, serta ketiadaan umpan balik visual langsung (*lack of immediate feedback loop*) mereduksi rasa empati dan rasa bersalah pelaku di balik sekat layar gawai [3], [4].

Guna menekan laju penyebaran konten berbahaya tersebut, kehadiran sistem moderasi konten otomatis berbasis kecerdasan buatan menjadi kebutuhan yang sangat mendesak. Pendekatan moderasi manual maupun sistem konvensional berbasis pencocokan kata kunci (*keyword matching*) telah kehilangan efektivitas operasionalnya di tengah ledakan volume data. Komunikasi warganet di media sosial memiliki karakteristik linguistik yang sangat

volatil, tidak terstruktur, serta sarat akan penggunaan bahasa gaul (*slang*), singkatan tidak baku, pencampuran kode bahasa (*code-mixing*), kesalahan pengetikan (*typo*), hingga manipulasi tipografi yang sengaja dirancang untuk menghindari filter kata [5], [6]. Di samping kompleksitas struktural tersebut, moderasi konten modern tidak lagi memadai jika hanya mengadopsi klasifikasi biner (menentukan teks berbahaya atau aman secara dikotomi). Sistem dituntut mampu memetakan muatan teks ke dalam spektrum multi-kelas yang mencerminkan lima tingkatan keparahan dampak (Aman, Ringan, Sedang, Tinggi, dan Sangat Berbahaya) guna menentukan skala prioritas penanganan dan mitigasi risiko secara terukur dan presisi [7], [8].

Dalam ranah *Natural Language Processing* (NLP), keandalan model klasifikasi teks multi-kelas bertumpu pada efektivitas teknik representasi kata (*word embedding*). Metode *static word embedding* seperti FastText memiliki keunggulan komparatif dalam menangani kata baru di luar kosakata korpus latih (*Out-of-Vocabulary* / OOV) karena memanfaatkan representasi *character n-grams* [9]. Pendekatan berbasis subkata ini memungkinkan FastText merekonstruksi vektor kata tidak baku atau istilah *slang* berdasarkan kemiripan morfologis komponen penyusunnya. Namun, representasi vektor FastText bersifat tetap (*fixed vector*) dan mengabaikan urutan kata dalam kalimat. Keterbatasan ini membuat model statis gagal menyelesaikan masalah polisemi (kata bermakna ganda) serta tidak mampu mengidentifikasi pergeseran makna semantik akibat kehadiran kata negasi, ekspresi sarkasme, maupun konteks kalimat tersirat [10], [11].

Sebagai terobosan mutakhir, arsitektur Transformer [12] melahirkan model bahasa kontekstual seperti IndoBERT yang telah dilatih secara ekstensif pada korpus bahasa Indonesia berskala miliaran kata [13], [14]. Mekanisme *self-attention* pada IndoBERT memungkinkan ekstraksi konteks dua arah secara mendalam (*deep bidirectional context*) yang dinamis mengikuti interaksi kata dengan struktur kalimat di sekelilingnya. Vektor yang dihasilkan IndoBERT sangat adaptif dalam membedakan nuansa teks yang ambigu pada interaksi media sosial. Kendati demikian, keunggulan pemahaman kontekstual IndoBERT menghadirkan konsekuensi berupa kebutuhan sumber daya komputasi yang berat (*resource-intensive*) dan waktu pelatihan yang jauh lebih lama dibandingkan metode statis.

Tantangan metodologis lainnya terletak pada karakteristik FastText sebagai ekstraktor fitur pasif yang tidak memiliki mekanisme inferensi keputusan mandiri [15], [16]. Oleh karena itu, guna mengekstrak batas kinerja optimal dari representasi vektor statis FastText secara objektif sebelum dikomparasikan dengan IndoBERT, fitur tersebut perlu diuji bersama dua algoritma pembelajaran mesin eksternal dengan karakteristik geometri batas keputusan (*decision boundary*) yang berbeda. *Logistic Regression* diterapkan untuk mewakili model linear parametrik yang efisien dan stabil [17], sedangkan *Random Forest* digunakan untuk mewakili model ansambel non-linear berbasis pohon keputusan yang tangguh menangani data berdimensi kompleks. Hasil terbaik dari kombinasi FastText tersebut selanjutnya diuji tanding secara *head-to-head* melawan kepala klasifikasi internal (*Dense Layer* dan fungsi *Softmax*) milik IndoBERT.

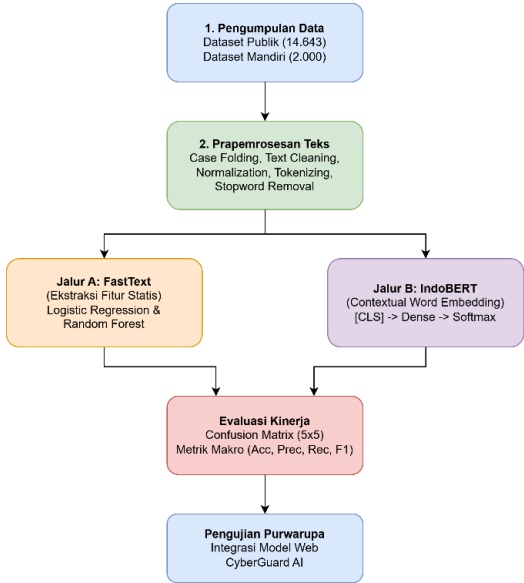
Meskipun kajian deteksi perundungan siber terus berkembang, studi-studi terdahulu masih menyisakan sejumlah keterbatasan analitis. Pertama, penelitian Ibrohim dan Budi (2019) terkait deteksi ujaran kebencian multi-label masih bertumpu pada metode ekstraksi fitur konvensional (TF-IDF) yang terbukti rentan terhadap pergeseran makna akibat kalimat sarkasme. Kedua, meskipun tolok ukur IndoLEM oleh Koto et al. (2020) telah memperkenalkan keandalan model bahasa IndoBERT, pengujiannya mayoritas masih terbatas pada ranah analisis sentimen biner dan pengenalan emosi dasar. Ketiga, eksplorasi deteksi tingkat keparahan (*severity*) perundungan oleh Ali et al. (2020) menggunakan algoritma pembelajaran mesin belum membedah secara komprehensif ketangguhan arsitektur Transformer pada korpus teks informal yang bergejolak. Keempat, penerapan deep learning untuk memonitor perilaku agresif di media sosial oleh Han et al. (2024) belum secara spesifik mengevaluasi perbandingan performa antara arsitektur static subword dengan contextual embedding pada kondisi ketidakseimbangan data (*class imbalance*) yang tajam. Berdasarkan keterbatasan dari keempat studi tersebut, terlihat jelas adanya kesenjangan penelitian (*research gap*) yang mendasar: belum ada studi komparatif komprehensif yang secara eksplisit membedah dan menguji efektivitas FastText versus IndoBERT dalam skenario klasifikasi teks imbalanced multi-class pada lima skala keparahan (Aman, Ringan, Sedang, Tinggi, dan Sangat Berbahaya) pada korpus komentar berbahasa Indonesia. Menjawab kesenjangan tersebut, kontribusi utama penelitian ini berfokus pada dua metrik esensial: (1) memberikan analisis empiris head-to-head untuk memecahkan masalah polisemi dan leksikon bermutasi pada spektrum klasifikasi keparahan yang kompleks, dan (2) menghadirkan kebaruan praktis dengan merealisasikan model pemrosesan bahasa alami terbaik ke dalam purwarupa aplikasi web interaktif bernama "CyberGuard AI" guna memfasilitasi moderasi konten waktu-nyata.

2. METODOLOGI PENELITIAN

2.1 Desain Penelitian dan Alur Kerja

Penelitian ini menerapkan jenis penelitian komparatif (*comparative research*) dengan pendekatan eksperimental untuk menganalisis dan membandingkan efektivitas metode *static subword embedding* (FastText) dan *contextual word embedding* (IndoBERT) dalam mendeteksi lima tingkatan keparahan *cyberbullying* pada komentar media sosial berbahasa Indonesia [19], [20]. Secara komprehensif, tahapan penelitian diuraikan ke dalam diagram alur yang

mencakup pengumpulan data, prapemrosesan, pembagian jalur eksperimen, hingga inferensi model dan pengujian purwarupa, sebagaimana divisualisasikan pada Gambar 1.



Gambar 1. Alur Kerja

2.2 Korpus Data Penelitian dan Skema Pelabelan

Korpus yang digunakan dalam penelitian ini berjumlah 16.643 data teks komentar berbahasa Indonesia yang merepresentasikan dinamika komunikasi digital kontemporer. Korpus dibangun melalui integrasi dataset publik dan data hasil pengumpulan mandiri (*crawling*):

Dataset Publik (14.643 sampel): Diperoleh dari repositori sekunder teranotasi, mencakup *Dataset Indonesian Hate Speech and Abusive Language* Twitter sebanyak 13.169 teks dan *Dataset Klasifikasi Ujaran Kebencian* sebanyak 1.476 teks. Dataset Mandiri (2.000 sampel): Diperoleh melalui penambahan data otomatis memanfaatkan skrip Python (*Tweepy*, *snsrape*, dan *Instaloader*) dari platform Twitter/X (1.263 data) dan Instagram (737 data).

Data hasil penambahan mandiri kemudian dianotasi secara manual (*self-annotation*) ke dalam lima tingkatan skala keparahan *cyberbullying* berdasarkan muatan semantis dan potensi dampak psikososialnya. Kriteria dan distribusi data untuk masing-masing kelas pelabelan direpresentasikan secara rinci pada Tabel 1.

Tabel 1. Skala Keparahan *Cyberbullying*

Kode	Kategori	Karakteristik & Indikasi Muatan Semantik	Sampel (n)	Proporsi (%)
0	Aman	Teks netral, informatif, atau bermuatan apresiasi/dukungan positif.	11.738	70,53%
1	Ringan	Sindiran halus, penggunaan sarkasme, atau ejekan tidak langsung.	1.072	6,44%
2	Sedang	Ujaran peyoratif eksplisit, makian verbal kasar, dan body shaming.	357	2,15%
3	Tinggi	Pelecehan verbal berat dan ujaran kebencian berbasis identitas primordial (SARA).	2.747	6,51%
4	Sangat Berbahaya	Provokasi anarkis, ancaman kekerasan fisik, atau instruksi pembunuhan.	729	4,38%

Berdasarkan Tabel 1, distribusi data pada korpus mengindikasikan fenomena ketidakseimbangan antarkelas (*class imbalance*). Kategori Aman (0) mendominasi sebesar 70,53%, berbanding terbalik dengan kategori Sedang (2) yang memiliki proporsi terendah (2,15%). Kondisi ini menjadi landasan digunakannya metrik macro-averaging agar evaluasi kinerja model tidak terbias oleh kelas mayoritas. Lebih lanjut, sebagai upaya untuk meminimalisasi bias pelabelan dari sumber dataset gabungan, penelitian ini melakukan anotasi silang (*cross-validation annotation*) secara acak terhadap 10% data untuk menjamin validitas dan konsistensi pelabelan kelima tingkat keparahan.

2.3 Prapemrosesan Teks (*Text Preprocessing*)

Untuk mereduksi derau (*noise*) dan menyeragamkan variasi tipografi pada bahasa media sosial, teks masukan dialirkan melalui pipa alir prapemrosesan bertahap yang terdiri atas: *Case Folding* (menyeragamkan huruf kecil), *Text Cleaning* (penghapusan tautan, angka, tanda baca menggunakan *regex*), *Text Normalization* (konversi *slang* dan salah tik ke kata baku), *Tokenizing* (pemotongan kalimat menjadi larik kata), dan *Stopword Removal* (eliminasi kata hubung yang tidak berbobot).

2.4 Jalur Eksperimen Pemodelan

Eksperimen klasifikasi multi-kelas dijalankan melalui dua jalur arsitektur representasi teks yang dibandingkan kinerjanya.

Jalur Ekstraksi Fitur Statis (FastText)

FastText merupakan algoritma representasi kata yang memanfaatkan subword n-grams, sehingga model ini tangguh dalam menangani kata-kata baru yang tidak terdapat dalam kosakata korpus latih (Out-of-Vocabulary). Berdasarkan pendekatan subkata ini, sebuah vektor representasi kata utuh, yang disimbolkan dengan v_w , dibentuk melalui hasil agregasi atau penjumlahan dari seluruh representasi n-gram karakter penyusunnya. Secara matematis, kalkulasi ini didefinisikan pada Persamaan (1), di mana variabel G_w merepresentasikan himpunan n-gram karakter yang diekstrak dari suatu kata w , dan variabel z_g merupakan representasi vektor spasial untuk setiap unit subword n-gram g yang berhimpun di dalamnya.

$$v_w = \sum_{g \in G_w} z_g \quad (1)$$

Matriks fitur keluaran dari arsitektur FastText tersebut kemudian dieksekusi oleh algoritma Logistic Regression menggunakan pendekatan fungsi aktivasi Multinomial Softmax [Tambahkan Sitasi Buku/Jurnal Machine Learning] dengan persamaan probabilitas yang ditunjukkan pada Persamaan (2). Dalam persamaan tersebut, $\sigma(z_i)$ menyatakan probabilitas prediksi akhir untuk kelas ke- i . Nilai ini diperoleh dengan mengkalkulasikan eksponensial dari nilai logit kelas ke- i (disimbolkan dengan e^{z_i}), yang kemudian dibagi dengan total jumlah eksponensial dari seluruh kelas yang ada, di mana K merupakan representasi jumlah total kelas klasifikasi (dalam penelitian ini $K=5$).

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad (2)$$

Algoritma pembandingan kedua pada jalur ini adalah Random Forest, yang merupakan model ansambel berbasis pohon keputusan dengan konfigurasi 15 trees dan skema majority voting.

Jalur Representasi Kontekstual (IndoBERT)

Menggunakan model pre-trained *indobert-base-uncased*. Vektor penanda token [CLS] pada lapisan akhir (final hidden state) dihubungkan ke lapisan linear tambahan (*Dense Layer*) dan dinormalisasi menggunakan fungsi *Softmax*. Hyperparameter yang digunakan: *learning rate* 2×10^{-5} , *batch size* 16, fungsi optimasi AdamW, fungsi kerugian *Cross-Entropy Loss*, dan dilatih selama 5 *epochs*.

2.5 Evaluasi Kinerja dan Metrik Pengukuran

Evaluasi kinerja model diukur menggunakan instrumen *Confusion Matrix* berdimensi 5×5 untuk memetakan luaran *True Positive* (TP_k), *True Negative* (TN_k), *False Positive* (FP_k), dan *False Negative* (FN_k). Guna mengatasi potensi bias akibat ketidakseimbangan proporsi data (*class imbalance*), perhitungan metrik dilakukan menggunakan pendekatan rata-rata makro (*Macro-Averaging*) dengan $K = 5$:

Accuracy (Akurasi Total):

$$\text{Accuracy} = \frac{\sum_{k=1}^K TP_k}{\text{Total Data Uji}} \quad (3)$$

Macro Precision

$$\text{Precision}_{\text{macro}} = \frac{1}{K} \sum_{k=1}^K \frac{TP_k}{TP_k + FP_k} \quad (4)$$

Macro Recall:

$$\text{Recall}_{\text{macro}} = \frac{1}{K} \sum_{k=1}^K \frac{TP_k}{TP_k + FN_k} \quad (5)$$

Macro F1-Score:

$$\text{F1-Score}_{\text{macro}} = \frac{1}{K} \sum_{k=1}^K \frac{2 \times \text{Precision}_k \times \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k} \quad (6)$$

Untuk merepresentasikan performa model secara komprehensif, Persamaan (3) hingga (6) menggunakan beberapa variabel evaluasi utama. Variabel K menetapkan batas jumlah kelas klasifikasi yang digunakan ($K=5$), dengan i sebagai indeks penunjuk kelas saat ini (dimulai dari kelas 0 hingga 4). Luaran klasifikasi diukur melalui empat metrik dasar: TP_i (True Positive) yang menunjukkan jumlah data kelas i yang berhasil diprediksi secara tepat sebagai kelas i ; TN_i (True Negative) yang mempresentasikan jumlah data bukan kelas i dan diprediksi dengan tepat sebagai bukan kelas i . Sebaliknya, instabilitas prediksi model ditandai oleh FP_i (False Positive) untuk data bukan kelas i yang secara keliru ditebak sebagai kelas i , serta FN_i (False Negative) yang merekam jumlah data kelas i

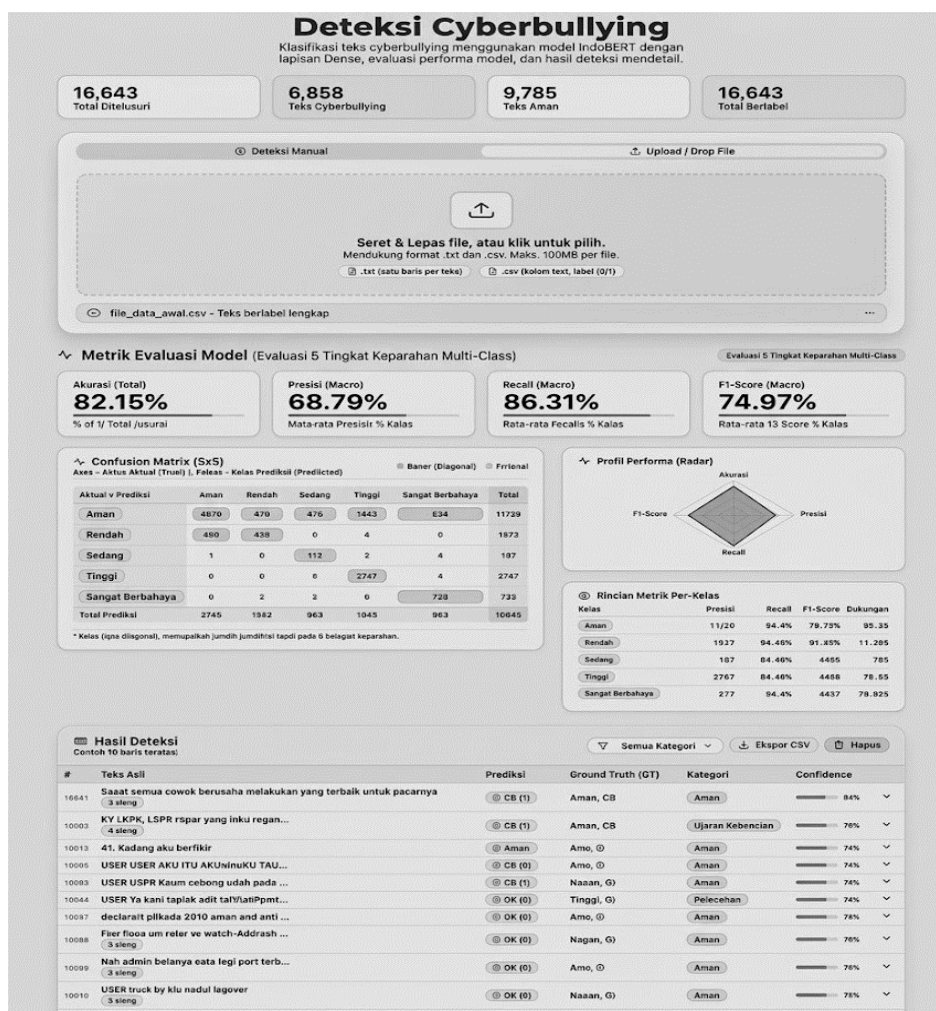
namun luput terprediksi karena diklasifikasikan sebagai kelas lain. Model dengan nilai metrik makro (F1-Score) tertinggi kemudian diintegrasikan ke dalam antarmuka purwarupa aplikasi web interaktif bernama CyberGuard AI yang dibangun berbasis framework Flask/FastAPI untuk memfasilitasi deteksi teks secara waktu-nyata (real-time) maupun inferensi dokumen kolektif (batch processing).

3. HASIL DAN PEMBAHASAN

Hasil penelitian ini berisi studi komparatif metode IndoBERT dan FastText, dimana metode IndoBERT dengan Algoritma Dense Layer dan metode FastText terdiri dari Algoritma Logistic Regresion dan Random Forest dengan membandingkan Akurasi, Presisi, Recall, dan F1-Score untuk masing-masing metode. Untuk mempermudah pengguna, sistem diimplementasikan ke dalam antarmuka aplikasi web.

3.1 Implementasi Antarmuka Sistem

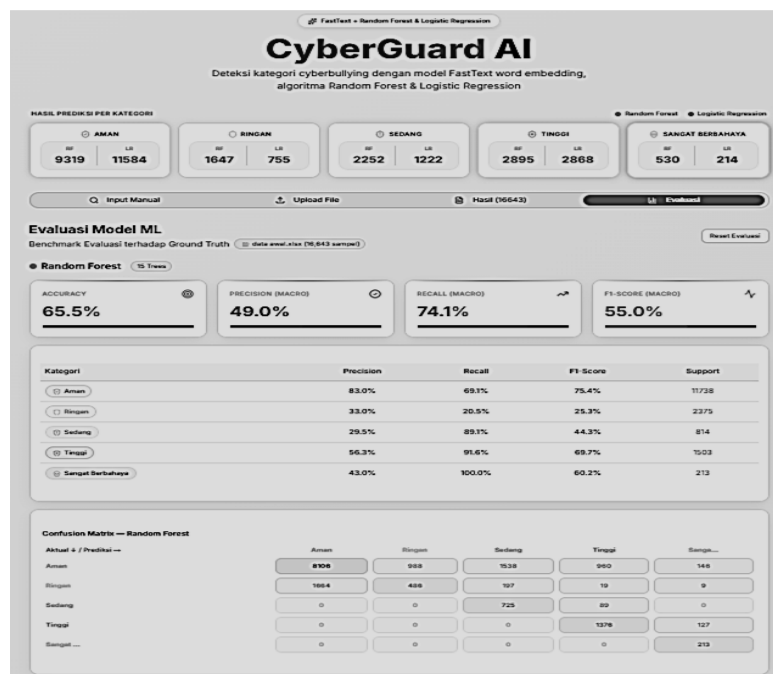
Untuk membuktikan efektivitas implementasi model klasifikasi dalam skenario moderasi konten secara nyata, hasil pelatihan dan evaluasi diintegrasikan ke dalam sebuah purwarupa aplikasi berbasis web mandiri (*standalone web application*) bernama "CyberGuard AI". Antarmuka sistem dirancang untuk memfasilitasi pengujian inferensi waktu-nyata (*real-time*), pemrosesan berkas masif (*batch processing*), serta visualisasi metrik evaluasi komparatif secara interaktif. Visualisasi antarmuka implementasi model IndoBERT direpresentasikan pada Gambar 1. Selanjutnya, evaluasi model FastText bersama algoritma Random Forest ditunjukkan pada Gambar 2, dan hasil evaluasi FastText bersama Logistic Regression didemonstrasikan pada Gambar 3.



Gambar 2. Antar Muka IndoBERT dengan Algoritma Dense Layer

Antarmuka web deteksi *cyberbullying* ini menampilkan implementasi model IndoBERT dengan *Dense Layer* secara interaktif. Bagian atas dasbor menyajikan kartu ringkasan total 16.643 data yang dianalisis, pembagian biner antara 6.858 teks *cyberbullying* dan 9.785 teks *non-cyberbullying*, serta modul pengujian teks manual dan fitur unggah berkas. Pada panel evaluasi, ditampilkan empat metrik utama rata-rata makro Akurasi Total 82,15%, Presisi Makro 68,79%, Recall Makro 86,31%, dan F1-Score Makro 74,97% yang dilengkapi dengan grafik radar dan tabel *Confusion Matrix*.

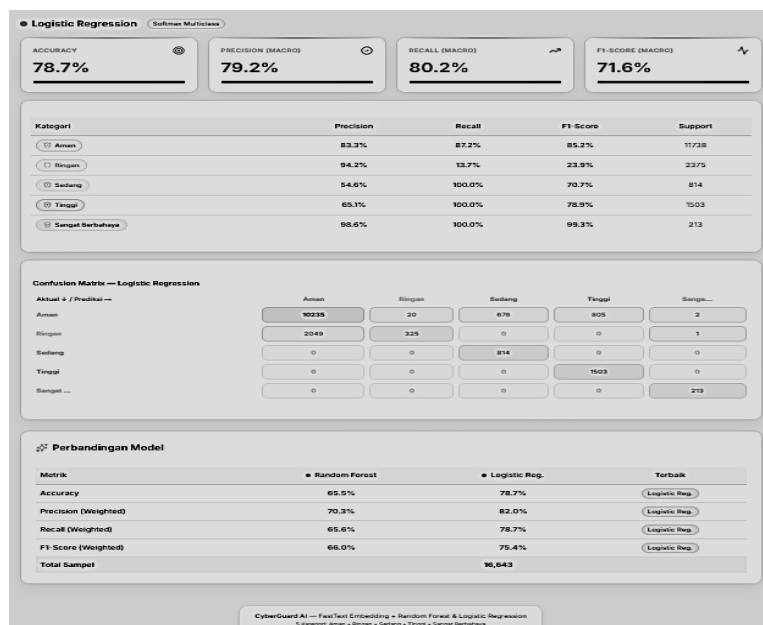
Bagian bawah antarmuka memuat tabel rekam jejak hasil deteksi secara rinci untuk setiap teks komentar. Tabel ini menyajikan teks asli beserta rencana deteksi jumlah kata *slang*, verifikasi prediksi terhadap label acuan (*ground truth*), klasifikasi lima spektrum keparahan (Aman, Ringan, Sedang, Tinggi, dan Sangat Berbahaya), serta bilah persentase tingkat keyakinan (*confidence score*) berbasis aktivasi *Softmax*, lengkap dengan tombol filter kategori dan ekspor berkas CSV.



Gambar 3. Antar Muka FastText dengan Algoritma Random Forest dan Logistic Regression 1

Tampilan antarmuka aplikasi CyberGuard AI ini menyajikan hasil implementasi metode ekstraksi fitur *static word embedding* FastText yang dieksekusi menggunakan algoritma ansambel *Random Forest* (15 trees). Pada bagian atas halaman, sistem menampilkan lima kartu ringkasan kategori keparahan yang membandingkan perolehan prediksi antara *Random Forest* (RF) dan *Logistic Regression* (LR), disertai bilah navigasi menu pengujian teks (*Input Manual*, *Upload File*, *Hasil*, dan *Evaluasi*).

Pada panel evaluasi model *Random Forest*, antarmuka menampilkan empat metrik makro utama, yaitu Akurasi sebesar 65,5%, Presisi Makro 49,0%, Recall Makro 74,1%, dan F1-Score Makro 55,0% terhadap 16.643 sampel data uji. Bagian tengah dan bawah halaman dilengkapi tabel rincian metrik per kategori (*Precision*, *Recall*, *F1-Score*, dan *Support*) serta visualisasi *Confusion Matrix* yang memetakan persebaran klasifikasi benar pada diagonal utama dan mendokumentasikan letak kesalahan prediksi model secara transparan.



Gambar 4. Antar Muka FastText dengan Algoritma Random Forest dan Logistic Regression 2

Antarmuka aplikasi web CyberGuard AI pada halaman ini menampilkan hasil evaluasi model linier *Logistic Regression (Softmax Multiclass)* yang dipadukan dengan representasi kata FastText. Bagian atas dasbor memvisualisasikan empat kartu metrik evaluasi makro, yaitu Akurasi sebesar 78,7%, Presisi Makro 79,2%, Recall Makro 80,2%, dan F1-Score Makro 71,6%. Tepat di bawahnya, disajikan tabel rincian metrik per kategori yang mendemonstrasikan tingkat *Recall* sempurna (100%) pada kelas Sedang, Tinggi, dan Sangat Berbahaya, meskipun tingkat *Recall* pada kelas Ringan mengalami penurunan hingga 13,7%.

Bagian tengah dan bawah halaman memuat matriks konfusi serta modul ringkasan perbandingan model. *Confusion Matrix* memperlihatkan keberhasilan klasifikasi pada sumbu diagonal (seperti 10.235 data Aman dan 1.503 data Tinggi), sekaligus mencatat 2.049 sampel data kategori Ringan yang salah terprediksi ke kategori Aman. Pada panel terbawah, tabel "Perbandingan Model" secara eksplisit menyandingkan metrik rata-rata terbobot (*Weighted-Average*) antara *Random Forest* dan *Logistic Regression* pada 16.643 sampel data, yang menegaskan kinerja yang lebih superior *Logistic Regression* sebagai algoritma pendamping FastText terbaik di seluruh parameter uji.

3.2 Hasil Model IndoBERT + Dense Layer

Pengujian model kontekstual IndoBERT yang dipadukan dengan kepala klasifikasi *Dense Layer* dan fungsi aktivasi *Softmax* terhadap keseluruhan 16.643 data teks komentar media sosial berbahasa Indonesia menghasilkan pemetaan performa yang konsisten pada skema lima tingkat keparahan. Evaluasi granular berbasis *Confusion Matrix* berdimensi 5x5 disajikan secara komprehensif pada Tabel 2.

Tabel 2. Confusion Matrix IndoBERT + Dense Layer (5×5 Multi-Class)

Aktual \ Prediksi	Aman (0)	Ringan (1)	Sedang (2)	Tinggi (3)	Sangat Berbahaya (4)	Total
Aman (0)	9.276	479	306	1.443	234	11.738
Ringan (1)	509	563	0	0	0	1.072
Sedang (2)	0	0	357	0	0	357
Tinggi (3)	0	0	0	2.747	0	2.747
Sangat Berbahaya (4)	0	0	0	0	729	729
Total Prediksi	9.785	1.042	663	4.190	963	16.643

Berdasarkan sebaran matriks pada Tabel 2, akumulasi elemen diagonal utama yang merepresentasikan klasifikasi benar (*True Positive*) mencapai 13.672 sampel dari total 16.643 data uji. Berdasarkan formula pengukuran matematis, model IndoBERT berhasil mengamankan nilai Akurasi Total sebesar 82,15%, Presisi Makro (*Macro Precision*) 68,79%, Recall Makro (*Macro Recall*) 86,31%, dan F1-Score Makro (*Macro F1-Score*) sebesar 74,97%.

Secara analitis, ketangguhan arsitektur Transformer pada model IndoBERT terlihat sangat dominan pada spektrum perundungan tingkat menengah hingga kritis. Rincian metrik per-kelas membuktikan bahwa IndoBERT mencatatkan tingkat sensitivitas (*Recall*) sempurna sebesar 100,00% pada kategori Sedang (357 dari 357 data), Tinggi (2.747 dari 2.747 data), dan Sangat Berbahaya (729 dari 729 data). Perolehan *Recall* absolut ini menegaskan bahwa tidak ada satu pun (*zero False Negative*) konten perundungan berat, ujaran kebencian berbasis identitas primordial (SARA), maupun ancaman kekerasan fisik yang lolos tanpa terdeteksi oleh sistem. Kapabilitas ini berakar dari mekanisme *self-attention* dua arah (*deep bidirectional context*) yang mampu menyerap representasi kontekstual secara mendalam melalui token khusus [CLS], sehingga model sangat sensitif terhadap leksikon peyoratif eksplisit maupun konstruksi kalimat intimidatif.

Kendati demikian, matriks konfusi mencatat adanya tantangan klasifikasi pada batas perbatasan (*borderline cases*) antara kelas Aman dan Ringan. Sebanyak 509 sampel teks perundungan kategori Ringan secara keliru diprediksi ke dalam kategori Aman, dan 479 teks kategori Aman terprediksi ke kategori Ringan. Pola kesalahan ini dipicu oleh karakteristik linguistik perundungan ringan di media sosial yang kerap menggunakan gaya bahasa sindiran halus, sarkasme ambigu, atau ekspresi candaan antarteman yang leksikonnya sangat menyerupai tuturan percakapan kasual normal. Selain itu, sebanyak 1.443 teks aman bergeser ke kategori Tinggi akibat kehadiran kata-kata kasar yang dilontarkan tanpa target agresif personal, namun tetap direspons secara waspada oleh bobot atensi model. Meskipun terdapat pergeseran pada tingkat ringan, konsistensi model dalam mempertahankan batas keputusan (*decision boundary*) yang rasional menjadikan kombinasi IndoBERT dan *Dense Layer* sebagai solusi paling andal dan protektif untuk moderasi konten digital secara otomatis.

3.3 Hasil Model FastText + Random Forest & Logistic Regression

Pengujian jalur ekstraksi fitur static word embedding FastText memperlihatkan perbedaan performa yang sangat kontras antara algoritma non-linear Random Forest (RF) dan model linear Logistic Regression (LR) pada skema evaluasi 5 tingkat keparahan. Rincian komparasi metrik kinerja per-kelas dari kedua algoritma inferensi tersebut disajikan pada Tabel 3.

Tabel 3. Rincian Evaluasi Metrik Per-Kelas: FastText + RF vs FastText + LR

Kategori Keparahan	Support (N)	FastText + Random Forest			FastText + Logistic Regression		
		Presisi	Recall	F1-Score	Presisi	Recall	F1-Score
Aman (0)	11.738	83,0%	69,1%	75,4%	83,3%	87,2%	85,2%

Kategori Keparahan	Support (N)	FastText + Random Forest			FastText + Logistic Regression		
		Presisi	Recall	F1-Score	Presisi	Recall	F1-Score
Ringan (1)	2.375	33,0%	20,5%	25,3%	94,2%	13,7%	23,9%
Sedang (2)	814	29,5%	89,1%	44,3%	54,6%	100,0%	70,7%
Tinggi (3)	1.503	56,3%	91,6%	69,7%	65,1%	100,0%	78,9%
Sangat Berbahaya (4)	213	43,0%	100,0%	60,2%	98,6%	100,0%	99,3%
Rata-rata Makro	16.643	49,0%	74,1%	55,0%	79,2%	80,2%	71,6%
Akurasi Total	—	65,5% (10.906 / 16.643 sampel)			78,7% (13.264 / 16.643 sampel)		

Seperti yang diuraikan pada Tabel 3, algoritma Logistic Regression membuktikan keunggulan yang konsisten di hampir seluruh parameter pengukuran, mencatatkan Akurasi Total sebesar 78,7% (13.264 sampel benar) dan F1-Score Makro 71,6%. Model linear ini sangat efektif menangani spektrum keparahan tinggi dengan perolehan Recall mencapai 100,0% pada kategori Sedang, Tinggi, dan Sangat Berbahaya. Kendati demikian, model ini mengalami kendala sensitivitas pada kategori Ringan (Recall 13,7%) akibat tingginya angka salah klasifikasi ke kategori Aman (2.049 sampel meleset). Tingginya tingkat False Positive pada prediksi kelas Aman ini mengindikasikan bahwa representasi statis FastText kehilangan detail semantik saat menjumpai kalimat yang secara struktur tampak netral namun menyisipkan satu leksikon bermuatan negatif, di mana model merespons bobot kata secara linier tanpa mampu memahami konteks kalimat secara utuh.

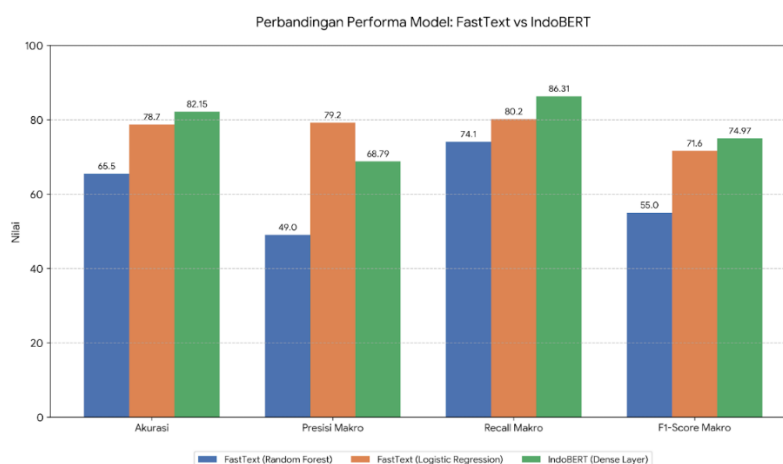
Sebaliknya, pengujian Random Forest menghasilkan penurunan performa yang signifikan dengan Akurasi Total terendah sebesar 65,5% (10.906 sampel benar), Presisi Makro 49,0%, dan F1-Score Makro 55,0%. Penurunan ini dipicu oleh kegagalan partisi ortogonal pada pohon keputusan ansambel dalam menarik batas pemisah di ruang vektor kontinu berdimensi padat. Temuan ini membuktikan secara empiris bahwa model parametrik linear jauh lebih kompatibel dan andal untuk mendampingi metode static subword embedding FastText dibandingkan arsitektur ansambel berbasis pohon keputusan non-linear.

3.4 Komparasi Kinerja Antar-Model

Perbandingan performa komparatif secara menyeluruh antara metode *static subword embedding* (FastText) dan *contextual word embedding* (IndoBERT) dengan berbagai kombinasi algoritma klasifikasi dirangkum pada Tabel 4.

Tabel 4. Perbandingan Performa Komparatif Keseluruhan Model

Metode Representasi	Algoritma Klasifikasi			Akurasi (%)	Presisi Makro (%)	Recall Makro (%)	F1-Score Makro (%)
FastText (Static Subword)	Random Forest	(15 Trees)		65,50%	49,00%	74,10%	55,00%
FastText (Static Subword)	Logistic Regression			78,70%	79,20%	80,20%	71,60%
IndoBERT (Contextual)	Dense Layer + Softmax			82,15%	68,79%	86,31%	74,97%



Gambar 5. Grafik Komparasi Performa Keseluruhan Model

Guna memberikan representasi visual yang lebih intuitif terhadap disparitas kinerja antarmodel, komparasi metrik evaluasi (Akurasi, Presisi Makro, Recall Makro, dan F1-Score Makro) direpresentasikan melalui Gambar 5. Visualisasi grafik batang tersebut memperlihatkan secara jelas dominasi arsitektur kontekstual IndoBERT yang secara konsisten mengungguli kedua varian algoritma representasi statis FastText pada hampir seluruh parameter pengujian, menjadikannya arsitektur yang paling optimal untuk klasifikasi multi-kelas.

Berdasarkan rekapitulasi data pada Tabel 4, model kontekstual IndoBERT yang dipadukan dengan *Dense Layer* dan fungsi aktivasi Softmax membuktikan keunggulan komparatif mutlak atas seluruh varian model

representasi statis. IndoBERT berhasil mengamankan nilai Akurasi Total tertinggi sebesar 82,15%, Recall Makro sebesar 86,31%, dan F1-Score Makro mencapai 74,97%.

Di sisi lain, pengujian representasi kata statis FastText memperlihatkan disparitas performa yang sangat tajam. Kombinasi FastText dengan algoritma linear *Logistic Regression* mampu menghasilkan performa yang relatif kompetitif dan stabil dengan Akurasi 78,70%, Presisi Makro 79,20%, dan F1-Score Makro 71,60%. Sebaliknya, perpaduan FastText dengan algoritma ansambel non-linear *Random Forest* mencatatkan kinerja terendah di seluruh parameter evaluasi. Temuan komparatif ini membuktikan secara empiris bahwa arsitektur kontekstual IndoBERT merupakan pendekatan paling tangguh, rasional, dan akurat untuk menyelesaikan tugas deteksi perundungan siber multi-kelas pada komentar media sosial berbahasa Indonesia.

3.5 Pembahasan

Secara sintesis, hasil komparasi empiris dalam penelitian ini menegaskan sejumlah keunggulan ilmiah mendalam dan kebaruan praktis yang memberikan kontribusi nyata bagi pengembangan literatur moderasi konten otomatis berbahasa Indonesia. Berbeda secara mendasar dengan mayoritas studi terdahulu yang umumnya membatasi cakupan analisis pada deteksi dikotomi biner (hanya melabeli teks sebagai rundung atau tidak rundung), penelitian ini membuktikan keberhasilan perancangan sistem yang mampu memetakan muatan teks ke dalam lima tingkatan skala keparahan dampak secara granular, yaitu Aman, Ringan, Sedang, Tinggi, dan Sangat Berbahaya. Keunggulan pemetaan multi-kelas ini memberikan landasan taktis yang sangat presisi bagi pengembang platform digital, penyedia layanan interaksi virtual, maupun pemangku kebijakan untuk menerapkan eskalasi mitigasi risiko secara proporsional dan terukur sesuai dengan tingkat bahaya pesan yang terdeteksi.

Keunggulan distingtif utama dari integrasi arsitektur IndoBERT dengan lapisan klasifikasi Dense Layer bertumpu pada kemampuannya menarik batas keputusan (decision boundary) yang sangat rasional berbasis pemahaman semantik kontekstual yang mendalam. Model tidak terjebak dalam perangkat bias pencocokan kata kotor (keyword matching) konvensional yang kerap memicu tingginya angka False Positive. Secara cerdas, IndoBERT mampu membedakan antara tuturan vulgar non-personal yang tidak memiliki target korban dengan ekspresi intimidasi terstruktur yang memang ditujukan untuk merendahkan martabat individu tertentu. Di saat yang sama, model kontekstual ini menjamin keandalan proteksi ruang digital secara maksimal melalui perolehan nilai sensitivitas (Recall) mutlak sebesar 100% pada spektrum perundungan tingkat Sedang, Tinggi, dan Sangat Berbahaya. Capaian ini memastikan bahwa tidak ada konten bernada pelecehan berat, ujaran kebencian berbasis identitas primordial (SARA), maupun ancaman kekerasan fisik nyata yang lolos dari penyaringan sistem.

Sebagai penegasan kontribusi ilmiah, capaian kinerja optimal dari model IndoBERT dalam penelitian ini (Akurasi 82,15% dan Macro F1-Score 74,97%) menunjukkan keunggulan kompetitif dan peningkatan kapabilitas analisis yang signifikan jika disandingkan dengan penelitian-penelitian terdahulu. Sebagai komparasi langsung, penelitian Ibrahim dan Budi (2019) yang mengembangkan deteksi abusive language menggunakan ekstraksi fitur konvensional TF-IDF dan algoritma Support Vector Machine (SVM) hanya mampu mencapai nilai F1-Score pada kisaran 73% hingga 77%, itupun terbatas pada klasifikasi biner dan kerap gagal mendeteksi konteks sarkasme. Penelitian ini membuktikan bahwa arsitektur IndoBERT mampu mengungguli batasan metrik tersebut, bahkan pada tantangan klasifikasi lima tingkatan keparahan (multi-class) yang beban komputasi dan kompleksitas semantiknya jauh lebih tinggi. Selanjutnya, jika dibandingkan dengan tolok ukur IndoLEM oleh Koto et al. (2020) yang mencatatkan F1-Score di atas 80% namun distribusinya dibatasi pada analisis sentimen biner dan emosi dasar, penelitian ini berhasil memperluas spektrum implementasi model bahasa kontekstual tersebut. Model IndoBERT dalam penelitian ini tidak hanya mencetak Akurasi keseluruhan yang tangguh (82,15%), tetapi juga secara spesifik berhasil mengamankan tingkat sensitivitas (Recall) absolut sebesar 100% pada kelas perundungan ekstrem (kategori Sedang, Tinggi, dan Sangat Berbahaya). Pencapaian identifikasi mutlak pada batas keputusan (decision boundary) kelas minoritas ini merupakan terobosan yang sulit direplikasi oleh penelitian sebelumnya yang hanya mengandalkan pendekatan statis. Hal ini secara definitif mengonfirmasi bahwa pendekatan contextual embedding tidak sekadar unggul secara komparasi angka, tetapi menawarkan ketahanan semantik komprehensif yang melampaui temuan-temuan pada studi moderasi teks sebelumnya.

Dari sudut pandang teoretis pemrosesan bahasa alami, penelitian ini memberikan kontribusi ilmiah yang berharga dengan membedah secara eksplisit perilaku geometris ruang representasi kata ketika diintegrasikan dengan berbagai paradigma algoritma inferensi. Temuan empiris membuktikan bahwa ruang matriks vektor padat kontinu hasil ekstraksi subword FastText memiliki kompatibilitas dan keterikatan linear yang kuat saat dipadukan dengan pemisah linear parametrik Logistic Regression, sehingga mampu menghasilkan performa klasifikasi yang relatif stabil dan terhindar dari bias ekstrem. Sebaliknya, pengujian ini membuktikan bahwa arsitektur ansambel berbasis partisi pohon keputusan non-linear (Random Forest) mengalami kegagalan fundamental dalam menarik batas pemisah pada ruang vektor padat FastText, yang berujung pada fragmentasi fitur dan bias klasifikasi masif, terutama pada kategori perundungan tingkat Ringan.

Nilai lebih yang menjadi pembeda utama dari penelitian ini terletak pada kesiapan implementasinya di dunia nyata (real-world readiness). Penelitian ini tidak berhenti pada eksperimen komputasi algoritma dan penalaan parameter di lingkungan tertutup, melainkan berhasil mengonversi seluruh arsitektur pipa alir pemrosesan teks—mulai dari modul pembersihan karakter, normalisasi kata gaul (slang) berbasis kamus leksikal, ekstraksi representasi semantik, hingga kalkulasi probabilitas keyakinan Softmax—ke dalam purwarupa aplikasi web interaktif bernama

"CyberGuard AI". Sistem ini dirancang adaptif untuk memfasilitasi pengujian inferensi kalimat tunggal secara waktu-nyata (real-time) maupun analisis dokumen korpus berskala besar (batch processing). Sinergi antara keandalan pemahaman semantik Transformer dan arsitektur sistem perangkat lunak yang modular ini menegaskan bahwa pendekatan contextual embedding dinamis merupakan fondasi paling tangguh, protektif, dan akuntabel untuk mengatasi volatilitas bahasa media sosial yang sarat akan dialek gaul, sarkasme, dan ambiguitas makna.

4. KESIMPULAN

Penelitian ini membuktikan bahwa representasi contextual word embedding (IndoBERT) memiliki keunggulan komparatif signifikan dibandingkan static subword embedding (FastText) dalam tugas klasifikasi cyberbullying multi-kelas pada bahasa Indonesia informal, dengan pencapaian Akurasi 82,15% dan Macro F1-Score 74,97%. Kemampuan IndoBERT menyerap konteks dua arah terbukti esensial dalam meminimalisasi False Negative pada spektrum perundungan ekstrem (Tinggi dan Sangat Berbahaya) dibandingkan pendekatan statis berbasis Logistic Regression dan Random Forest. Meskipun purwarupa sistem telah menunjukkan kelayakan inferensi waktu-nyata, keterbatasan penelitian ini terletak pada tingginya instabilitas identifikasi pada perundungan kategori Ringan (sarkasme/sindiran laten) serta ukuran sampel pada kelas minoritas. Untuk pengembangan akademis dan praktis selanjutnya, direkomendasikan penggunaan teknik data augmentation untuk mengatasi ketidakseimbangan kelas minoritas, serta pengujian arsitektur hibrida yang menggabungkan fitur linguistik sintaktis bersama Transformer untuk meningkatkan sensitivitas pada batas keputusan (decision boundary) kategori marginal. Sebagai kontribusi akhir, penelitian ini memberikan landasan taktis dan empiris bagi pengembang platform digital bahwa arsitektur kontekstual dinamis merupakan instrumen yang jauh lebih presisi, unggul, dan protektif dalam memitigasi eskalasi dampak destruktif perilaku cyberbullying berbahasa Indonesia.

5. DEKLARASI PENULIS

5.1 Kontribusi Penulis

Konseptualisasi: M. I. A.;

Metodologi: M. I. A.;

Perangkat Lunak: M. I. A.;

Validasi: M. I. A.;

Analisis Formal: M. I. A.;

Investigasi: M. I. A.;

Sumber Daya: M. I. A.;

Kurasi Data: M. I. A.;

Penulisan Draf Awal: M. I. A.;

Penulisan, Peninjauan, dan Penyuntingan: M. I. A.;

Visualisasi: M. I. A.;

Seluruh penulis telah membaca dan menyetujui versi manuskrip yang telah diterbitkan.

5.2 Ketersediaan Data

Data yang disajikan dalam penelitian ini tersedia berdasarkan permintaan kepada penulis pertama.

5.3 Pendanaan

Penelitian ini tidak menerima pendanaan eksternal.

5.4 Pernyataan Dewan Peninjau Institusional

Tidak Berlaku (Not applicable)

5.5 Pernyataan Persetujuan Berdasarkan Informasi

Tidak Berlaku (Not applicable)

5.6 Pernyataan Konflik Kepentingan

Para penulis menyatakan bahwa mereka tidak memiliki kepentingan keuangan yang bersaing atau hubungan pribadi yang diketahui yang dapat dianggap memengaruhi pekerjaan yang dilaporkan dalam artikel ini.

5.7 Penggunaan AI Generatif dan Teknologi Berbantuan AI dalam Proses Penulisan

Para penulis menggunakan AI Generatif untuk meningkatkan kejelasan penulisan dalam artikel ini. Para penulis telah meninjau dan menyunting konten yang dihasilkan dengan bantuan AI serta bertanggung jawab penuh atas versi akhir publikasi.

REFERENCES

- [1] D. Riswanto and R. Marsinun, "Perilaku Cyberbullying Remaja di Media Sosial," *ANALITIKA*, vol. 12, no. 2, pp. 98–111, 2020, doi: 10.31289/analitika.v12i2.3704.
- [2] A. N. A. W, C. Santika, I. Wahyudi, and K. M. Istiqomah, "Perundungan Dunia Maya Menjadi Faktor Utama Depresi Berujung Bunuh Diri," *De Juncto Delicti Journal of Law*, vol. 4, no. 2, pp. 15–29, 2024, doi: 10.35706/djd.v4i2.7972.
- [3] V. Auriemma, G. Iorio, G. Roberti, and R. Morese, "Cyberbullying and Empathy in the Age of Hyperconnection: An Interdisciplinary Approach," *Frontiers in Sociology*, vol. 5, p. 551881, 2020, doi: 10.3389/fsoc.2020.551881.
- [4] M. D. Sánchez-Hernández, M. C. Herrera, and F. Expósito, "Is Online Disinhibition Related to Cyberdating Abuse Perpetration through Moral Disengagement? The Moderating Role of Gender, Sexism, and Cybervictimization," *Sex Roles*, vol. 90, no. 7, pp. 938–959, 2024, doi: 10.1007/s11199-024-01480-3.
- [5] S. Govindankutty and A. Kumar, "Design and Implementation of Automated Content Moderation Systems in Social Media," *Integrated Journal for Research in Arts and Humanities*, vol. 4, no. 6, pp. 380–402, 2024, doi: 10.55544/ijrah.4.6.27.
- [6] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," pp. 46–57, 2019, doi: 10.18653/v1/w19-3506.
- [7] H. Han, M. Asif, E. M. Awwad, N. Sarhan, Y. Y. Ghadi, and B. Xu, "Innovative deep learning techniques for monitoring aggressive behavior in social media posts," *Journal of Cloud Computing Advances Systems and Applications*, vol. 13, no. 1, 2024, doi: 10.1186/s13677-023-00577-6.
- [8] A. Muzakir and K. Adi, "Classification of Hate Speech Language Detection on Social Media: Preliminary Study for Improvement," vol. 4, pp. 1–11.
- [9] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," vol. 5, pp. 135–146, 2020.
- [10] J. Devlin, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," no. Mlm, 2019.
- [11] A. Onan and M. A. Toçoğlu, "A Term Weighted Neural Language Model and Stacked Bidirectional LSTM Based Framework for Sarcasm Identification," *IEEE Access*, vol. 9, pp. 7701–7722, 2021, doi: 10.1109/access.2021.3049734.
- [12] A. Vaswani, "Attention Is All You Need," no. Nips, 2017.
- [13] B. Wilie *et al.*, *IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding*. 2020. doi: 10.48550/arXiv.2009.05387.
- [14] F. Koto and T. Baldwin, "IndoLEM and IndoBERT : A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," pp. 757–770, 2020.
- [15] A. Joulin, "Bag of Tricks for Efficient Text Classification," vol. 2, pp. 427–431, 2017.
- [16] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning Word Vectors for 157 Languages," Feb. 2018, doi: 10.48550/arXiv.1802.06893.
- [17] M. Ghosh and G. Sanyal, "An ensemble approach to stabilize the features for multi-domain sentiment analysis using supervised machine learning," *J. Big Data*, vol. 5, no. 1, 2018, doi: 10.1186/s40537-018-0152-5.
- [18] B. Ali, T. Id, and D. O. Sullivan, "Cyberbullying severity detection : A machine learning approach," pp. 1–19, 2020, doi: 10.1371/journal.pone.0240924.
- [19] D. Purwitasari, A. S. S. Ansyah, A. P. Kurniawan, and A. N. Kholifah, "A Hybrid Method on Emotion Detection for Indonesian Tweets of COVID-19," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 2, pp. 254–262, 2023, doi: 10.29207/resti.v7i2.4816.
- [20] G. G. Setiaji, J. Suntoro, and A. Rifa'i, "Random State Parameter Undersampling untuk Penanganan Data dengan Kelas Tidak Seimbang pada Algoritme Random Forest," *Jurnal Transformatika*, vol. 21, no. 2, pp. 73–83, 2024, doi: 10.26623/transformatika.v21i2.8901.