Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9249

Deteksi Cyberbullying pada Komentar Media Sosial Berbahasa Indonesia Menggunakan Pendekatan Hibrida IndoBERTweet- BiLSTM

Reza Ramadhon Aditya*, Arry Maulana Syarif

Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Dian Nuswantoro, Semarang, Indonesia
Email: 1,*111202214467@mhs.dinus.ac.id, 2, arry.maulana@dsn.dinus.ac.id
Email Penulis Korespondensi: 111202214467@mhs.dinus.ac.id
Submitted: **19/01/2026**; Accepted: **05/03/2026**; Published**: 06/03/2026**

Abstrak−Cyberbullying pada media sosial berbahasa Indonesia merupakan permasalahan serius yang berdampak signifikan terhadap kesehatan mental dan kualitas interaksi digital. Karakteristik bahasa media sosial yang informal, ambigu, serta kaya akan slang dan sarkasme menjadi tantangan utama dalam pengembangan sistem deteksi otomatis yang akurat. Penelitian ini mengusulkan pendekatan hibrida untuk deteksi cyberbullying dengan mengintegrasikan model bahasa pra-latih spesifik domain IndoBERTweet dan arsitektur Bidirectional Long Short-Term Memory (BiLSTM). IndoBERTweet dimanfaatkan untuk menghasilkan representasi semantik kontekstual yang selaras dengan karakteristik bahasa Twitter Indonesia, sementara BiLSTM digunakan untuk menangkap ketergantungan sekuensial dua arah pada tingkat kalimat. Eksperimen dilakukan menggunakan dataset publik Twitter berbahasa Indonesia yang telah dianotasi secara manual, dengan total 13.091 data yang dikonversi ke dalam skema klasifikasi biner. Untuk mengatasi ketidakseimbangan kelas, diterapkan kombinasi strategi class weighting dan label smoothing selama proses pelatihan. Kinerja model dievaluasi menggunakan metrik Accuracy, Precision, Recall, F1-Score, ROC-AUC, dan PR-AUC. Hasil eksperimen menunjukkan bahwa model IndoBERTweet–BiLSTM mencapai performa terbaik dengan perolehan F1-Score sebesar 87,53%, Recall 88,80%, Precision 86,31%, serta skor ROC-AUC 92,91% dan PR-AUC 94,25%. Capaian ini secara konsisten mengungguli model baseline berbasis IndoBERT maupun IndoBERT-p1 dengan konfigurasi arsitektur yang identik. Temuan ini menegaskan pentingnya keselarasan domain (domain alignment) dalam meningkatkan akurasi deteksi cyberbullying pada teks media sosial berbahasa Indonesia.

Kata Kunci: Cyberbullying; Deteksi Teks; IndoBERTweet; BiLSTM; Media Sosial; NLP Bahasa Indonesia

Abstract−Cyberbullying on Indonesian-language social media has become a serious issue with significant psychological and social consequences, necessitating the development of reliable automated detection systems. However, the informal, ambiguous, and highly contextual nature of social media language, including the frequent use of slang and sarcasm, poses substantial challenges for conventional text classification approaches. This study proposes a hybrid cyberbullying detection model that integrates the domain-specific pre-trained language model IndoBERTweet with a Bidirectional Long Short-Term Memory (BiLSTM) architecture. IndoBERTweet is employed to generate contextualized semantic representations aligned with the linguistic characteristics of Indonesian Twitter data, while BiLSTM is utilized to capture bidirectional sequential dependencies at the sentence level. Experiments were conducted using a publicly available, manually annotated Indonesian Twitter dataset consisting of 13,091 samples, which were reformulated into a binary classification scheme. To address class imbalance, a combination of class weighting and label smoothing was applied during model training. Model performance was evaluated using Accuracy, Precision, Recall, F1-Score, ROC-AUC, and PR-AUC metrics. Experimental results show that the IndoBERTweet–BiLSTM model achieved the best performance with an F1-Score of 87.53%, Recall of 88.80%, Precision of 86.31%, ROC-AUC of 92.91%, and PR-AUC of 94.25%. This performance consistently outperforms baseline models based on IndoBERT and IndoBERT-p1 with identical architectural configurations. These findings highlight the critical role of domain alignment in enhancing cyberbullying detection performance for Indonesian social media text.

Keywords: Cyberbullying; Text Classification; IndoBERTweet; BiLSTM; Social Media; Indonesian NLP

1. PENDAHULUAN

Fenomena *cyberbullying* telah berkembang menjadi ancaman serius dalam ekosistem komunikasi digital, terutama di platform media sosial yang memfasilitasi anonimitas dan jangkauan masif [1], [2]. Cyberbullying didefinisikan sebagai perilaku agresif yang dilakukan secara sengaja dan berulang melalui media digital dengan tujuan menyakiti individu atau kelompok tertentu [1], [3]. Dampak psikologis dari paparan cyberbullying sangat signifikan, mencakup gangguan kesehatan mental, penurunan prestasi akademik, hingga peningkatan risiko depresi dan ideasi bunuh diri [4]. Di Indonesia, pertumbuhan eksponensial pengguna media sosial seperti Twitter dan TikTok memperparah permasalahan ini, terutama karena maraknya penggunaan bahasa informal, slang regional, dan ekspresi sarkastik yang bersifat kontekstual dalam interaksi daring [5], [6], [7].

Deteksi cyberbullying pada teks berbahasa Indonesia menghadapi tantangan linguistik yang kompleks dan multidimensional [5], [7]. Tantangan pertama adalah ketidakbakuan struktur bahasa, di mana pengguna media sosial sering mengabaikan kaidah tata bahasa formal, menggunakan singkatan non-standar seperti "gue", "elu", dan "wkwk", serta melakukan repetisi karakter untuk penekanan emosional seperti "anjiiing" atau "bangettt" [6], [8], [9]. Tantangan kedua adalah ambiguitas kontekstual yang sangat tinggi, kata-kata kasar yang secara leksikal bersifat negatif dapat digunakan dalam konteks keakraban antar teman tanpa intensi menyakiti, sementara serangan implisit melalui *sarkasme* dapat tidak mengandung kata kasar sama sekali [10], [11], [12]. Tantangan ketiga adalah dinamika bahasa yang cepat berubah, dengan munculnya slang baru dan meme secara konstan yang membuat pendekatan berbasis kamus leksikon statis menjadi cepat ketinggalan zaman [4], [5]. Kompleksitas linguistik ini menyebabkan pendekatan

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9249

moderasi manual tidak lagi scalable seiring dengan volume data yang terus meningkat, sehingga sistem deteksi otomatis berbasis pemrosesan bahasa alami menjadi kebutuhan yang mendesak [13].

Pendekatan pembelajaran mesin konvensional seperti *Support Vector Machine* dan *Random Forest* telah banyak digunakan dalam penelitian awal deteksi cyberbullying, namun metode-metode ini memiliki keterbatasan fundamental dalam menangkap representasi semantik mendalam dan memahami konteks kalimat yang kompleks [14], [15]. Perkembangan model bahasa berbasis arsitektur *Transformer*, khususnya BERT, membawa terobosan signifikan melalui mekanisme pemahaman konteks dua arah yang mampu menangkap hubungan antar kata dalam kalimat secara lebih komprehensif [7], [11]. Untuk bahasa Indonesia, IndoBERT telah menunjukkan performa yang menjanjikan dalam berbagai tugas pemrosesan bahasa alami [16]. Namun, kesenjangan domain menjadi hambatan kritis yang signifikan, IndoBERT dilatih menggunakan korpus teks formal seperti Wikipedia dan artikel berita, sehingga akurasinya cenderung menurun drastis ketika diaplikasikan pada teks media sosial yang didominasi bahasa informal dan struktur tidak teratur [5], [6].

Untuk mengatasi kesenjangan domain tersebut, beberapa penelitian menerapkan strategi *preprocessing* intensif seperti normalisasi slang berbasis leksikon untuk memetakan kata tidak baku ke bentuk formal [4], [8], [16]. Meskipun pendekatan ini dapat meningkatkan performa dalam batas tertentu, teknik berbasis kamus memiliki kelemahan inheren. Pertama, kamus leksikon bersifat statis dan tidak dapat mengikuti evolusi bahasa media sosial yang sangat dinamis, di mana slang baru dan variasi penulisan muncul hampir setiap hari [2], [6]. Kedua, normalisasi leksikal semata tidak mampu menangkap nuansa semantik implisit seperti sindiran atau intensi agresif yang tidak diekspresikan melalui kata-kata kasar secara eksplisit [9], [10], [11]. Keterbatasan fundamental ini mendorong urgensi penggunaan model bahasa pra-latih spesifik domain yang memiliki keselarasan natural dengan karakteristik teks media sosial, sehingga dapat memahami pola bahasa informal tanpa bergantung sepenuhnya pada normalisasi berbasis kamus [3], [5].

IndoBERTweet merupakan model bahasa pra-latih yang dikembangkan secara khusus menggunakan korpus Twitter berbahasa Indonesia dalam skala masif, mencakup 409 juta kata dari percakapan media sosial autentik [6], [17]. Karakteristik unik dari proses *pre-trained* memberikan model ini pemahaman intrinsik terhadap kosakata informal, pola sintaksis tidak baku, dan struktur kalimat khas media sosial yang jarang ditemukan dalam teks formal [5], [7]. Kesesuaian karakteristik antara data pre-trained dan data tugas hilir ini menciptakan *domain alignment* yaitu penyelarasan distribusi fitur yang meminimalkan kesenjangan antara *source domain* dan target domain yang terbukti mampu meningkatkan kualitas representasi semantik pada tugas klasifikasi teks informal, khususnya dalam menangani variasi bahasa tidak baku yang dominan pada konten cyberbullying. Di sisi lain, meskipun arsitektur Transformer unggul dalam menangkap dependensi jarak jauh melalui mekanisme *self-attention global*, integrasi dengan arsitektur sekuensial seperti *Bidirectional Long Short-Term Memory* masih relevan untuk memperkuat pemahaman ketergantungan urutan kata dari dua arah secara eksplisit [10], [11]. Kemampuan BiLSTM dalam memproses informasi sekuensial secara *forward* dan *backward* memberikan representasi kontekstual tambahan yang krusial untuk menganalisis struktur kalimat panjang dan kompleks yang sering muncul dalam konten cyberbullying [5], [15], [18].

Berdasarkan analisis kesenjangan penelitian terdahulu, penelitian ini mengusulkan pendekatan hibrida yang mengintegrasikan kekuatan representasi kontekstual spesifik domain dari IndoBERTweet dengan kemampuan pemodelan sekuensial BiLSTM untuk deteksi cyberbullying pada komentar media sosial berbahasa Indonesia [5], [6]. Dalam arsitektur yang diusulkan, mekanisme Global Max Pooling diterapkan pada lapisan pasca-BiLSTM untuk mereduksi dimensi fitur. Pemilihan strategi pooling ini didasarkan pada dua pertimbangan teknis utama: pertama, berbeda dengan Average Pooling yang cenderung mendilusi informasi, *Max Pooling* lebih efektif dalam menangkap fitur paling dominan seperti kata-kata toksik spesifik yang menjadi indikator kuat perilaku bullying. Kedua, pendekatan ini menawarkan efisiensi komputasi yang lebih baik dibandingkan penambahan lapisan Atensi, menjaga kompleksitas model tetap terkendali mengingat arsitektur BiLSTM sendiri sudah memiliki beban komputasi yang signifikan [10], [15], [18]. Untuk memastikan validitas temuan, penelitian ini menerapkan protokol eksperimental terkontrol dengan menggunakan konfigurasi hiperparameter yang identik untuk seluruh model yang diuji, sehingga perbedaan performa dapat diatribusikan secara langsung kepada perbedaan karakteristik model *pre-trained*, bukan kepada variasi *tuning*[3], [7], [11].

Secara spesifik, penelitian ini bertujuan untuk mengembangkan arsitektur hibrida IndoBERTweet-BiLSTM yang mengombinasikan representasi semantik spesifik domain dengan pemrosesan sekuensial dua arah untuk tugas deteksi cyberbullying berbahasa Indonesia [5], [6]. Penelitian ini juga bertujuan mengevaluasi secara empiris pengaruh domain alignment terhadap kinerja klasifikasi dengan membandingkan performa model berbasis domain media sosial terhadap model berbasis domain umum menggunakan konfigurasi arsitektur dan skema pelatihan yang setara dalam kondisi eksperimental terkontrol [3], [7]. Selain itu, penelitian ini menganalisis efektivitas strategi mitigasi ketidakseimbangan kelas melalui kombinasi *class weighting* dan *label smoothing* dalam konteks klasifikasi teks media sosial [14], [19], serta mengidentifikasi pola kesalahan klasifikasi untuk memahami batasan linguistik model dalam menangani ambiguitas kontekstual dan ujaran kebencian implisit [11], [20].

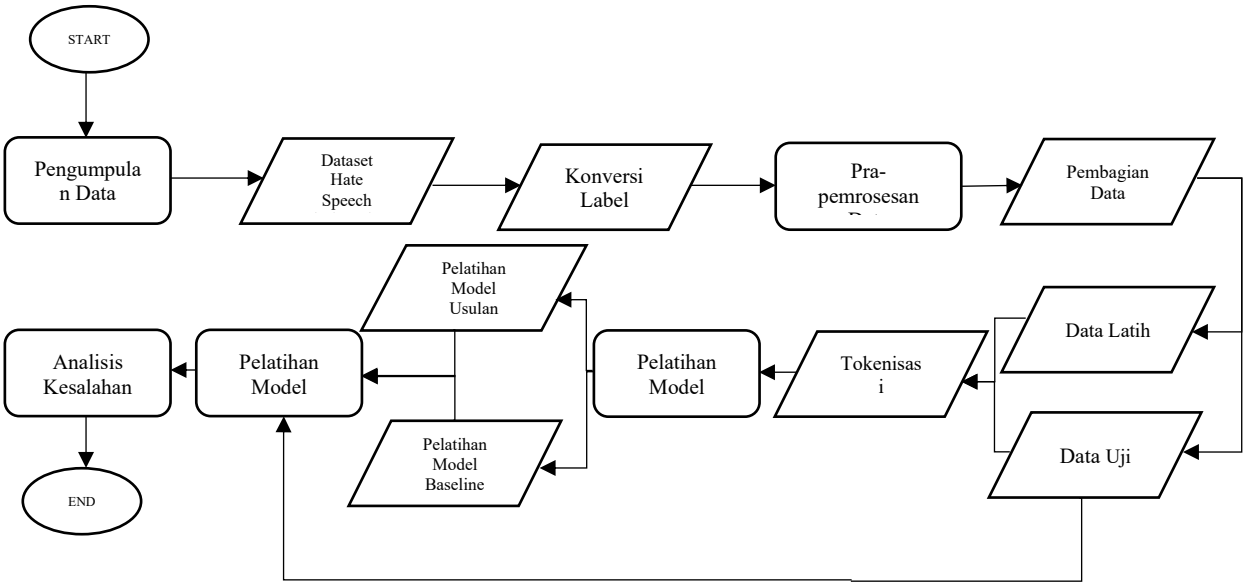
Penelitian ini memberikan tiga kontribusi utama. Pertama, penelitian ini mengusulkan kerangka deteksi cyberbullying berbasis pendekatan hibrida yang mengintegrasikan model bahasa pre-trained spesifik domain IndoBERTweet dengan arsitektur Bidirectional Long Short-Term Memory (BiLSTM) untuk menangkap representasi semantik kontekstual serta ketergantungan sekuensial dua arah pada teks media sosial berbahasa Indonesia. Kedua,

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9249

penelitian ini menyajikan analisis empiris terhadap pengaruh keselarasan domain (domain alignment) dengan membandingkan kinerja IndoBERTweet dan varian IndoBERT domain-umum menggunakan arsitektur serta konfigurasi pelatihan yang identik, sehingga evaluasi dilakukan secara adil dan terkontrol. Ketiga, penelitian ini melakukan analisis kesalahan untuk mengidentifikasi pola linguistik yang masih menantang bagi model berbasis transformer, seperti sarkasme, perundungan implisit, dan ambiguitas kontekstual, yang dapat menjadi landasan bagi pengembangan penelitian selanjutnya.

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan eksperimental kuantitatif dengan desain komparatif untuk mengevaluasi kinerja beberapa model klasifikasi berbasis Transformer pada tugas deteksi cyberbullying berbahasa Indonesia. Eksperimen dilakukan dengan membandingkan model bahasa pre-trained spesifik domain media sosial dan model berbasis domain umum menggunakan konfigurasi arsitektur dan skema pelatihan yang setara. Sistem penelitian dirancang mencakup alur pengumpulan dan pra-pemrosesan data, perancangan arsitektur model komparatif, pelatihan dengan strategi *handling class imbalance*, evaluasi komprehensif menggunakan metrik klasifikasi yang relevan, hingga analisis kesalahan untuk identifikasi pola misklasifikasi. Diagram sistem keseluruhan penelitian disajikan pada Gambar 1.



Gambar 1. *General System Penelitian*

2.1 Pengumpulan Data

Dataset yang digunakan dalam penelitian ini bersumber dari repositori publik [21], yang menghimpun data Twitter dan telah melalui proses anotasi manual oleh ahli bahasa untuk menjamin validitas labelnya. Korpus ini terdiri dari 13.170 tweet yang secara bawaan memiliki dua label terpisah: *Hate Speech* dan *Abusive*. Untuk memfokuskan penelitian pada deteksi cyberbullying secara menyeluruh, skema multi-label tersebut direstrukturisasi menjadi klasifikasi biner. Kelas target *Cyberbullying* ditentukan menggunakan logika disjungsi (*inclusive OR*), di mana sebuah tweet dikategorikan sebagai positif Cyberbullying jika mengandung unsur Hate Speech, Abusive atau keduanya. Sebaliknya, tweet dikategorikan sebagai *Non-Cyberbullying* hanya jika bersih dari kedua unsur tersebut. Dengan C, HS dan AB melambangkan Cyberbullying, Hate Speech dan Abusive, menggunakan rumus

$$C = HS \vee AB$$

atau,

$$C = \begin{cases} 1 \; jika \; HS = 1 \; atau \; AB = 1 \\ 0 \; jika \; HS = 0 \; dan \; AB = 0 \end{cases} \quad (1)$$

Dalam implementasi klasifikasi cyberbulling pada komenter media sosial berbahasa Indonesia, sistem akan mengklasifikasikan suatu komentar sebagai cyberbullying (C = 1) apabila terdeteksi mengandung salah satu atau kedua indikator HS dan AB bernilai 1, sementara konten yang bersih dari kedua unsur (HS = 0 dan AB = 0) akan diklasifikasikan sebagai Non-Cyberbullying (C = 0). Pendekatan ini memungkinkan deteksi yang lebih komprehensif dengan memanfaatkan logika OR sederhana, di mana keberadaan salah satu ciri bahaya sudah cukup untuk menandai konten sebagai berpotensi merugikan. Berdasarkan transformasi tersebut, diperoleh distribusi kelas yang terdiri dari 5.861 data kelas Non-Cyberbullying dan 7.309 data kelas Cyberbullying.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9249

Tabel 1 memperlihatkan contoh hasil transformasi label dari HS dan AB ke dalam C menggunakan rumus (1). Pada data sampel pertama, meskipun AB bernilai 0, label C ditetapkan bernilai 1 karena variabel HS bernilai 1. Pola serupa diterapkan pada data sampel kedua, di mana label C bernilai 1 dikarenakan variabel AB bernilai 1, dan HS bernilai 0. Sedangkan pada sampel data terakhir, label C ditetapkan bernilai 0 karena kedua variabel pembentuknya, HS dan AB, bernilai 0.

Tabel 1. Contoh hasil transformasi label dari Hate Speech dan Abusive ke dalam Cyberbullying

ID	Komentar	H	A	C
00001	- disaat semua cowok berusaha melacak perhatian gue. loe lantas remehkan perhatian yg gue kasih khusus ke elo. basic elo cowok bego ! ! !'	1	0	1
00002	RT USER: USER siapa yang telat ngasih tau elu?edan sarap gue bergaul dengan cigax jifla calis sama siapa noh licew juga'	0	1	1
…	…	…	…	…
13170	USER USER USER USER USER USER Yg jd pertanyaan : yg beli tolnya otomatis narik duit dari pengguna tol. Yg beli tolnya dari asing dan aseng. Berarti asing dan aseng ini sdh pake cara kuno bgitu.. ???'	0	0	0

2.2 Pra-Pemrosesan Data

Pra-pemrosesan data merupakan tahapan fundamental untuk mentransformasi teks mentah media sosial yang sarat dengan kebisingan (*noise*) menjadi representasi input yang optimal bagi model pembelajaran mendalam [5], [6]. Mengingat karakteristik bahasa media sosial yang tidak terstruktur serta variasi penulisan yang tinggi, rangkaian pra-pemrosesan diterapkan secara sekuensial dan konsisten. Hal ini bertujuan untuk mengoptimalkan konvergensi model serta mencegah terjadinya kebocoran data (*data leakage*) antara partisi latih dan uji [14].

Tahap awal pra-pemrosesan adalah pembersihan teks, yang bertujuan menghilangkan elemen non-semantik yang tidak berkontribusi terhadap makna kalimat [7]. Proses ini meliputi penghapusan tautan URL, *mention* (@username), hashtag, prefiks retweet (RT), serta karakter *non-ASCII* [3], [5]. Selain itu, dilakukan normalisasi repetisi karakter untuk mengurangi variasi penulisan berlebihan (misalnya “anjiiing” menjadi “anjing”). Selanjutnya, dilakukan normalisasi leksikon secara selektif berbasis kamus slang Indonesia untuk menangani variasi ortografi yang paling umum dan ekstrem dalam data media sosial. Pendekatan dictionary-based ini memungkinkan standardisasi terhadap variasi penulisan tertentu yang dapat mengganggu konsistensi tokenisasi seperti "gw" menjadi "gue", "loe" menjadi "kamu", atau "ngasih" menjadi "memberi" sambil secara strategis mempertahankan ribuan ekspresi slang informal lainnya yang tidak ter-cover dalam kamus, seperti kata-kata yang bersifat contextual, emerging slang, atau variasi regional. Meskipun model IndoBERTweet telah dilatih menggunakan korpus Twitter berbahasa Indonesia dan memiliki pemahaman intrinsik terhadap bahasa informal, normalisasi selektif ini diterapkan untuk menyeimbangkan antara reduksi noise leksikal ekstrem yang dapat menurunkan performa tokenizer dan preservasi karakteristik linguistik informal yang menjadi kekuatan model domain-specific. Strategi ini memastikan bahwa nuansa ekspresif bahasa media social yang esensial untuk deteksi cyberbullying tetap terjaga selama proses fine-tuning.

Setelah tahapan pembersihan dan normalisasi, dari total 13.170 tweet dalam dataset asli, terdapat 74 sampel (0,56%) yang menjadi string kosong akibat penghapusan elemen non-semantik (URL, mention, emoticon) sehingga tersisa 13.096 tweet. Selanjutnya, dilakukan filtering untuk menghilangkan sampel yang terlalu pendek (kurang dari 3 kata) guna memastikan kualitas data yang memadai untuk proses pembelajaran model, menghasilkan 13.091 sampel akhir yang digunakan dalam eksperimen. Total pengurangan sebesar 79 sampel (0,60%) dari dataset asli dianggap tidak signifikan terhadap distribusi kelas.

Tabel 2. Transformasi Teks - Sampel Teks Mentah vs Teks Bersih

Label	Teks Mentah	Teks Bersih (Cleaned)
1	- disaat semua cowok berusaha melacak perhatian gue. loe lantas remehkan perhatian yg gue kasih khusus ke elo. basic elo cowok bego ! ! !'	di saat semua cowok berusaha melacak perhatian gue. kamu lantas remehkan perhatian yang gue kasih khusus ke elo. basic kamu cowok bego
1	RT USER: USER siapa yang telat ngasih tau elu?edan sarap gue bergaul dengan cigax jifla calis sama siapa noh licew juga'	pengguna pengguna siapa yang telat memberi tau elu?edan sarap gue bergaul dengan cigax jifla calis sama siapa itu licew juga

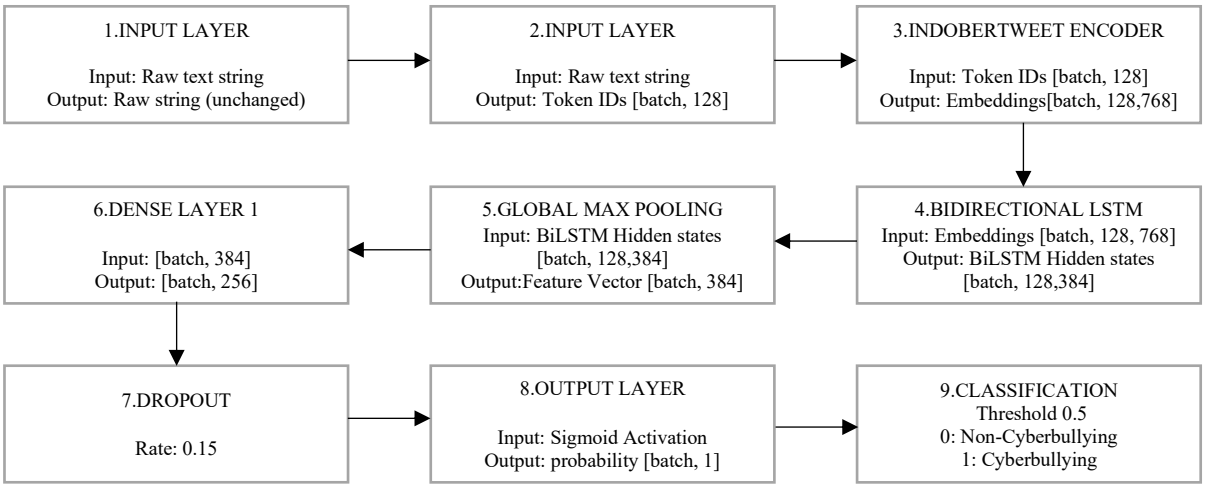
Representasi input bagi model dihasilkan melalui tokenisasi menggunakan tokenizer WordPiece terintegrasi IndoBERTweet, dengan penetapan panjang maksimum 128 token berdasarkan analisis distribusi panjang tweet yang menunjukkan bahwa lebih dari 95% sampel berada di bawah *threshold* tersebut setelah preprocessing. Strategi ini mengoptimalkan trade-off antara cakupan informasi semantik dan efisiensi komputasi, mengikuti protokol standar BERT untuk mempertahankan kompatibilitas dengan bobot pre-trained. Proses tokenisasi menghasilkan representasi *subword* yang konsisten untuk setiap tweet, dengan token khusus [CLS] sebagai penanda awal sekuens dan [SEP] sebagai penanda akhir, diikuti dengan *padding* hingga mencapai panjang maksimum yang telah ditetapkan. Contoh

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9249

transformasi teks dari tahap awal hingga hasil pra-pemrosesan disajikan pada Tabel 2, menunjukkan efektivitas pipeline pra-pemrosesan dalam menormalisasi variasi bahasa informal media sosial.

2.3 Perancangan Arsitektur Model

Penelitian ini mengusulkan arsitektur hibrida yang menggabungkan model pre-trained berbasis Transformer dengan Bidirectional Long Short-Term Memory (BiLSTM) dan lapisan Global Max Pooling. Arsitektur ini dirancang untuk memanfaatkan kemampuan representasi semantik dari IndoBERTweet dan kemampuan pemrosesan sekuensial dari BiLSTM tanpa menggunakan mekanisme attention tambahan untuk efisiensi komputasi. Detail arsitektur model diilustrasikan pada Gambar 2.



Gambar 2. *Arsitektur Model Usulan (IndoBERTweet-BiLSTM).*

Arsitektur hibrida yang diusulkan mengintegrasikan kekuatan representasi kontekstual spesifik-domain dari IndoBERTweet dengan kemampuan pemodelan dependensi sekuensial BiLSTM, sebagaimana diilustrasikan pada Gambar 2. Strategi ini didasarkan pada hipotesis bahwa kombinasi *contextual embedding* dari model pre-trained dengan *explicit sequential modeling* dapat menghasilkan representasi yang lebih robust untuk menangani kompleksitas linguistik cyberbullying, khususnya serangan implisit yang memerlukan pemahaman konteks kalimat secara holistik.

Pemilihan IndoBERTweet (indolem/indobertweet-base-uncased) sebagai *backbone* model didasarkan pada keselarasan domain (domain alignment) dengan karakteristik data target. Model ini dilatih pada korpus Twitter Indonesia berukuran masif (409 juta kata), sehingga hipotesis memiliki pemahaman intrinsik terhadap pola bahasa informal, struktur sintaksis tidak baku, dan kosakata slang yang dominan pada konten cyberbullying [17]. Keunggulan ini menjadi faktor kritis mengingat model pre-trained domain umum (IndoBERT) cenderung mengalami *performance degradation* ketika diaplikasikan pada teks media sosial tanpa fine-tuning substansial. Strategi fine-tuning selektif pada enam lapisan terakhir *encoder* diterapkan untuk mengoptimalkan *trade-off* antara adaptasi *task-specific* dan preservasi pengetahuan linguistik umum dari tahap pre-trained.

Integrasi lapisan BiLSTM dimotivasi oleh limitasi Transformer dalam menangkap dependensi sekuensial eksplisit untuk sekuens pendek-menengah seperti tweet [5], [18]. Meskipun self-attention mechanism mampu menangkap hubungan antar-token secara global, mekanisme rekurensi *bidirectional* memberikan *inductive bias* yang menguntungkan untuk memahami progresi intensi agresif dalam struktur kalimat [11].

Selanjutnya, Global Max Pooling diterapkan pada hidden states keluaran BiLSTM untuk mereduksi dimensi temporal menjadi representasi vektor tetap. Jika $H = \{h_1, h_2, \dots, h_T\}$ merepresentasikan urutan output BiLSTM dengan panjang sekuens T, maka operasi penyatuan (*pooling*) ini diformulasikan secara matematis dalam persamaan berikut:

$$v = max_{1 \le t \le T} h_t \tag{2}$$

Di mana v adalah vektor fitur hasil agregasi yang mengambil nilai maksimum dari setiap dimensi fitur di seluruh langkah waktu. Metode ini dipilih untuk efisiensi komputasi sambil mempertahankan kemampuan menangkap fitur-fitur paling menonjol (*salient features*) yang diskriminatif, menggantikan mekanisme attention tambahan yang cenderung meningkatkan kompleksitas model tanpa memberikan peningkatan performa proporsional pada eksperimen awal [5], [7].

Lapisan klasifikasi dirancang dengan regularisasi *Dropout* untuk mencegah *overfitting* pada dataset berukuran medium [23]. Fungsi aktivasi *Sigmoid* dipilih pada output layer karena kemampuannya mengkonversi nilai numerik arbitrary menjadi rentang probabilitas 0-1, yang sesuai dengan kebutuhan klasifikasi biner untuk memprediksi kemungkinan suatu tweet termasuk kategori cyberbullying atau non-cyberbullying. Output Sigmoid dapat diinterpretasikan langsung sebagai *confidence score* prediksi model terhadap kelas positif (cyberbullying). Arsitektur

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9249

keseluruhan mengikuti prinsip deep learning best practices dengan menyeimbangkan kapasitas model (model capacity) dan generalizability pada data cyberbullying yang memiliki karakteristik linguistik sangat variatif.

Lapisan klasifikasi dirancang dengan regularisasi Dropout untuk mencegah overfitting pada dataset berukuran medium, dengan fungsi aktivasi Sigmoid pada output layer untuk menghasilkan probabilitas klasifikasi biner [5]. Arsitektur keseluruhan mengikuti prinsip *deep learning best practi*ces dengan menyeimbangkan kapasitas model dan *generalizability* pada data cyberbullying yang memiliki karakteristik linguistik sangat variatif [3].

2.4 Pelatihan Model

Strategi pelatihan dirancang dengan mempertimbangkan karakteristik fine-tuning model Transformer pre-trained dan kompleksitas data cyberbullying yang memiliki ketidakseimbangan kelas moderat [7], [15]. Pembagian data dilakukan dengan memisahkan 20% data sebagai test set menggunakan *stratified splitting* untuk mempertahankan representasi proporsional kelas [6]. Dari 80% data training yang tersisa, 20% digunakan sebagai *validation set*, menghasilkan proporsi akhir 64% training, 16% validation, dan 20% testing dari total dataset, mengikuti protokol standar evaluasi pembelajaran mesin untuk memastikan validitas hasil pengujian [5].

Permasalahan class imbalance (rasio 1.25:1) ditangani melalui kombinasi *strategis class weighting* dan label smoothing, yang masing-masing mengatasi aspek berbeda dari ketimpangan data [19]. Class weighting memberikan penalti adaptif berdasarkan *invers frekuensi* kelas untuk mencegah bias prediksi terhadap kelas mayoritas, sementara label smoothing (α=0.05) berfungsi sebagai regularisasi implisit yang mencegah *overconfidence* model karakteristik penting untuk data cyberbullying yang *inherently* memiliki *boundary* ambiguitas tinggi antara ujaran kasar informal dan serangan personal [5], [6]. Pendekatan *dual-strategy* ini dipilih berdasarkan temuan literatur yang menunjukkan bahwa kombinasi keduanya menghasilkan performa superior dibanding penerapan tunggal pada tugas klasifikasi teks imbalanced [2], [3].

Konfigurasi hiperparameter mengikuti *best practices fine-tuning* Transformer dengan *learning rate* konservatif (3e-5) untuk mencegah catastrophic forgetting pada bobot pre-trained, *batch size* 16 untuk stabilitas konvergensi, dan regularisasi dropout moderat (0.15) untuk menjaga *trade-off generalization-capacity* [3]. Lapisan BiLSTM dikonfigurasi dengan 192 *hidden units* per arah (menghasilkan output 384-dimensional), diikuti *dense layer* 256 units untuk ekstraksi fitur sebelum klasifikasi akhir [5]. Jumlah *epoch* ditetapkan 10 iterasi penuh tanpa *early stopping* untuk menjaga konsistensi protokol eksperimental komparatif, didasarkan pada observasi bahwa model Transformer pre-trained umumnya mencapai konvergensi stabil dalam epoch rendah untuk tugas *downstream classification* [11]. *Optimizer Adam* digunakan dengan *gradient clipping* (norm=1.0) untuk mencegah *exploding gradients* selama fine-tuning [15].

Validasi metodologis dilakukan melalui perbandingan terkontrol dengan dua model baseline (IndoBERT-BiLSTM dan IndoBERT-P1-BiLSTM) menggunakan konfigurasi arsitektur dan hiperparameter identik [5], [6]. Uniformitas setup eksperimental ini krusial untuk memastikan bahwa perbedaan performa dapat diatribusikan secara langsung kepada perbedaan karakteristik model pre-trained (domain-specific vs. general-domain), bukan kepada variasi tuning atau arsitektur [3].

2.5 Evaluasi Model

Evaluasi kinerja model dilakukan menggunakan set data uji yang tidak pernah dilihat sebelumnya oleh model selama proses pelatihan. Mengingat karakteristik dataset cyberbullying yang sering kali memiliki distribusi kelas yang tidak seimbang, penelitian ini tidak hanya mengandalkan akurasi sebagai metrik utama. Evaluasi dilakukan secara komprehensif menggunakan metrik *Precision, Recall,* dan *F1-Score* untuk memberikan gambaran yang lebih objektif mengenai kemampuan model dalam mengklasifikasikan kelas minoritas.

Kinerja klasifikasi diukur berdasarkan nilai *True Positive* (TP), *False Positive* (FP), dan *False Negative* (FN) untuk menghitung metrik Precision, Recall, dan F1-Score. Precision (P) digunakan untuk mengukur tingkat keandalan model dalam memprediksi kelas positif, meminimalkan risiko kesalahan deteksi palsu. Metrik ini dihitung menggunakan persamaan berikut:

$$P = \frac{TP}{TP+FP} \tag{3}$$

Recall (R) mengukur sensitivitas model dalam mendeteksi seluruh insiden cyberbullying yang ada dalam dataset, dihitung menggunakan persamaan berikut:

$$R = \frac{TP}{TP+FN} \tag{4}$$

Metrik utama yang menjadi acuan adalah F1-Score, yang merupakan rata-rata harmonik antara Precision *dan* Recall. F1-Score memberikan gambaran performa yang lebih seimbang pada data yang memiliki distribusi kelas timpang, sebagaimana dinyatakan dalam persamaan berikut:

$$F1 = 2 \cdot \frac{P \cdot R}{P+R} \tag{5}$$

Selain metrik standar, penelitian ini menerapkan evaluasi berbasis threshold menggunakan *Area Under the Receiver Operating Characteristic Curve* (ROC-AUC) dan *Precision-Recall Area Under Curve* (PR-AUC). ROC-

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9249

AUC digunakan untuk mengukur daya pembeda model secara umum di berbagai ambang batas keputusan. Namun, mempertimbangkan distribusi kelas dataset yang sedikit tidak seimbang (*slightly imbalanced*), PR-AUC diadopsi sebagai metrik komplementer yang krusial. Berbeda dengan ROC-AUC yang dapat memberikan skor optimis pada data tidak seimbang, PR-AUC memberikan gambaran yang lebih reliabel mengenai performa model terhadap kelas *positif* (cyberbullying) dengan mengabaikan *True Negatives*, sehingga lebih sensitif dalam mengukur keseimbangan antara presisi dan recall[14].

2.6 Analisis Kesalahan

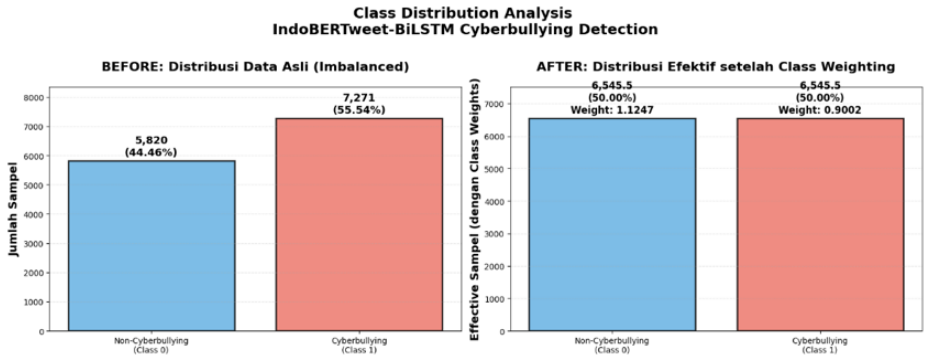
Selain evaluasi berbasis metrik kuantitatif, penelitian ini menerapkan analisis diagnostik mendalam menggunakan pendekatan Confusion Matrix untuk membedah pola kesalahan model. Pendekatan ini difokuskan pada pemetaan kasus FP dan FN guna mengidentifikasi kelemahan spesifik model dalam menangani ambiguitas bahasa yang sering memicu kesalahan klasifikasi [24], [25].Evaluasi kualitatif dilakukan secara manual terhadap sampel-sampel kesalahan tersebut untuk menelusuri fenomena linguistik yang menjadi penyebab kegagalan prediksi. Analisis ini bertujuan untuk memberikan interpretasi komprehensif mengenai batasan model hibrida dalam mendeteksi cyberbullying pada teks berbahasa Indonesia.

3. HASIL DAN PEMBAHASAN

Eksperimen deteksi cyberbullying menggunakan tiga varian model berbasis BERT yang dikombinasikan dengan BiLSTM menunjukkan hasil yang baik. Capaian ini didukung oleh evaluasi yang dilakukan secara komparatif antara model yang diusulkan (IndoBERTweet-BiLSTM) dengan dua model baseline (IndoBERT-BiLSTM dan IndoBERT-p1-BiLSTM) untuk menganalisis dampak keselarasan domain terhadap kinerja klasifikasi.

3.1 Statistik dan Distribusi Data

Setelah melalui tahapan pra-pemrosesan yang meliputi pembersihan dan normalisasi, total dataset yang digunakan dalam penelitian ini berjumlah 13.091 tweet. Analisis awal terhadap distribusi label menunjukkan karakteristik yang unik dibandingkan dataset pada umumnya, di mana kelas Cyberbullying (Label 1 atau C = 1) memiliki proporsi yang lebih dominan dibandingkan kelas Non-Cyberbullying (Label 0 atau C = 0). Gambar 3 (grafik kiri) memperlihatkan visualisasi kelas Cyberbullying yang berjumlah 7.271 sampel (55,54%), dan kelas Non-Cyberbullying yang berjumlah 5.820 sampel (44,46%). Kondisi ini mengindikasikan adanya ketidakseimbangan kelas (class imbalance) ringan dengan rasio sekitar 1,25:1. Dominasi konten negatif ini kemungkinan disebabkan oleh metode pengumpulan data yang menggunakan kata kunci spesifik terkait perundungan (keyword-based crawling) yang cenderung menjaring lebih banyak ujaran kebencian.

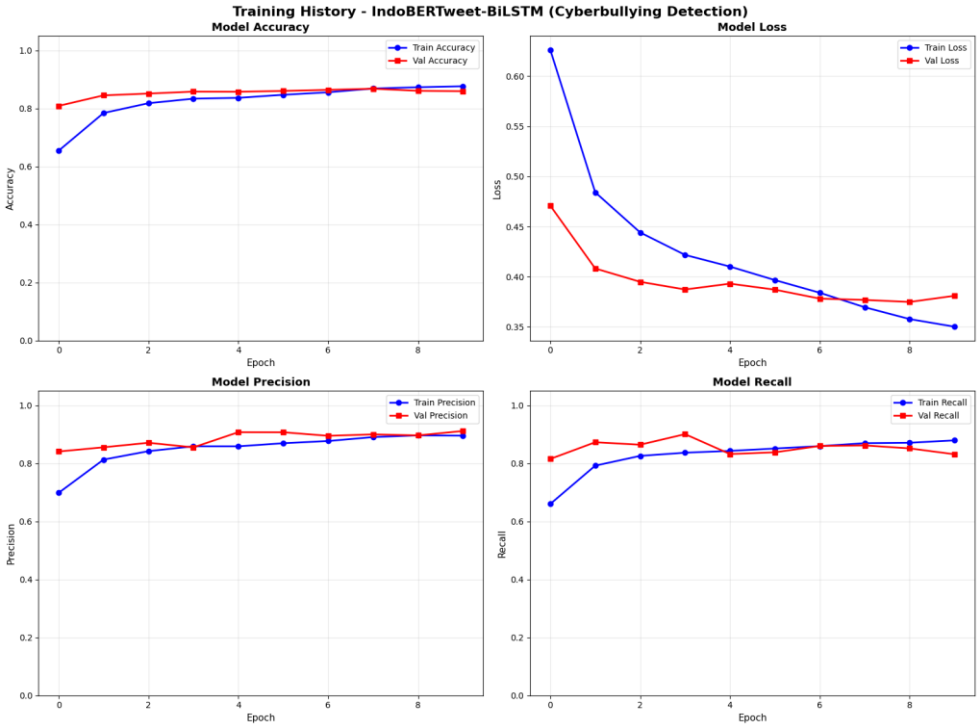


Gambar 3. *Distribusi Kelas Asli dan Visualisasi Efek Pembobotan Kelas (Class Weighting)*

Meskipun ketimpangan data tidak ekstrem, risiko bias model terhadap kelas mayoritas tetap ada. Untuk memitigasi hal ini tanpa mengurangi jumlah data fisik, penelitian ini menerapkan strategi Class Weighting. Teknik ini memberikan bobot penalti yang lebih besar pada kelas minoritas (Non-Cyberbullying) dan bobot lebih kecil pada kelas mayoritas. Gambar 2 (Grafik kanan), distribusi setelah class weighting, menunjukkan visualisasi dari distribusi efektif yang dipelajari oleh model selama proses pelatihan. Dengan penerapan bobot (Weight: 1.12 untuk kelas 0 dan Weight: 0.90 untuk kelas 1), model dipaksa untuk memperlakukan kedua kelas dengan prioritas yang setara (50%:50%), sehingga meminimalkan bias prediksi dan meningkatkan kemampuan generalisasi model pada kedua kelas.

Selain penerapan Class Weighting yang mengoptimalkan distribusi efektif data, penggunaan Label Smoothing turut berperan dalam stabilitas konvergensi model. Class Weighting berhasil memaksa model untuk lebih memperhatikan kelas minoritas, dan Label Smoothing juga efektif dalam menjaga agar bobot penalti yang besar tersebut tidak membuat model mengalami overfitting pada sampel-sampel tertentu yang bersifat ambigu atau memiliki noise tinggi. Sinergi kedua teknik ini terkonfirmasi pada grafik dinamika pelatihan (Gambar 3), di mana kurva loss bergerak stabil tanpa fluktuasi yang ekstrem.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9249



Gambar 4. *Grafik Dinamika Pelatihan IndoBERTweet-BiLSTM*

3.2 Hasil Pelatihan dan Analisis Konvergensi

Analisis mendalam terhadap dinamika proses pelatihan dilakukan untuk memverifikasi stabilitas model usulan (IndoBERTweet-BiLSTM) dan memastikan efektivitas strategi pembelajaran yang diterapkan. Model dilatih selama 10 epoch penuh tanpa penerapan early stopping guna memantau perilaku konvergensi secara menyeluruh. Ringkasan kinerja pelatihan pada tahap akhir disajikan pada Tabel 4, sementara visualisasi pergerakan metrik selama proses pembelajaran diilustrasikan pada Gambar 4.

Tabel 4. Ringkasan Metrik Pelatihan (Epoch 10)

Metrik Evaluasi	Data Latih (Train)	Data Validasi (Val)
Loss	0.3501	0.381
Accuracy	87.68%	85.97%
Precision	89.58%	91.19%
Recall	87.96%	83.16%
AUC Score	94.95%	94.20%

Berdasarkan riwayat pelatihan, terlihat pola konvergensi yang sangat stabil. Pada tahap awal (Epoch 1), model memulai dengan nilai loss sebesar 0,6262 dan berhasil menurunkannya secara signifikan hingga mencapai 0,3501 pada akhir epoch. Penurunan ini diikuti secara konsisten oleh kurva *Validation Loss* yang bergerak dari 0,4708 (Epoch 1) menjadi 0,3810 pada Epoch 10.

Tren kurva loss validasi yang tetap terkendali tanpa menunjukkan kenaikan ekstrem membuktikan bahwa mekanisme Label Smoothing sebesar 0,05 dan Dropout 0,15 bekerja efektif sebagai regulerisasi untuk mencegah overfitting. Indikator keberhasilan terlihat pada selisih (gap) antara akurasi data latih (87,68%) dan akurasi validasi (85,97%) yang relatif kecil, yaitu sebesar 1,71%. Hal ini mengonfirmasi bahwa model memiliki kemampuan generalisasi yang baik pada data yang belum pernah dilihat sebelumnya.

Meskipun terdapat sedikit fluktuasi pada metrik Recall validasi yang menyentuh angka 83,16% di akhir epoch, nilai Precision validasi justru meningkat hingga 91,19%. Hal ini menunjukkan bahwa sinergi antara Class Weighting dan Label Smoothing memungkinkan model untuk tetap belajar dari distribusi kelas yang telah diseimbangkan secara efektif. Dari sisi efisiensi, durasi pelatihan selama 10 epoch dinilai mencukupi bagi model berbasis Transformer untuk mencapai stabilitas performa.

3.3 Kinerja Model Usulan (IndoBERTweet-BiLSTM)

Evaluasi akhir model usulan dilakukan menggunakan data uji yang tidak terlibat dalam proses pelatihan maupun validasi. Berdasarkan hasil pengujian, model IndoBERTweet-BiLSTM menunjukkan performa yang sangat kuat dalam mendeteksi konten cyberbullying. Nilai ambang batas (threshold) optimal ditetapkan pada 0,35 untuk memaksimalkan kemampuan deteksi model. Ringkasan performa pada data uji disajikan dalam Tabel 5.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9249

Tabel 5 Metrik Evaluasi Kinerja Model Usulan pada Data Uji

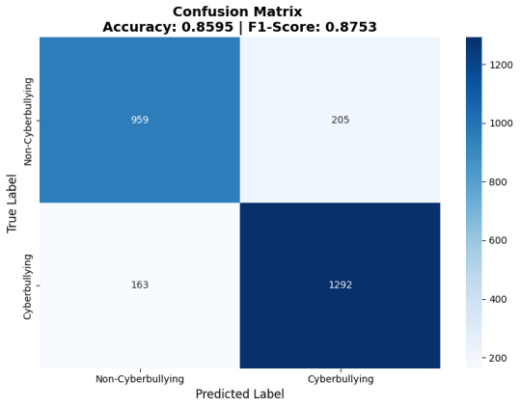
Metrik Evaluasi	Nilai
F1-Score (Kelas Cyberbullying)	0,8753
Recall (Kelas Cyberbullying)	0,8880
Precision (Kelas Cyberbullying)	0,8631
ROC-AUC Score	0,9291
PR-AUC Score	0,9425
Akurasi Keseluruhan	0,8595

Pada Tabel 5, terlihat bahwa model mampu mencapai nilai F1-Score sebesar 0,8753 untuk kelas cyberbullying. Tingginya nilai recall (0,8880) menunjukkan bahwa model sangat sensitif dalam menangkap indikasi perundungan di media sosial, yang merupakan aspek krusial dalam sistem moderasi konten.

Penetapan threshold pada 0,35 yang lebih rendah dari threshold default 0,5 merupakan keputusan strategis untuk memprioritaskan sensitivitas deteksi dalam konteks sistem moderasi konten. Threshold yang lebih rendah meningkatkan kemampuan model menangkap konten cyberbullying (recall 88,80%) dengan trade-off sedikit peningkatan false positive yang dapat direview manual oleh moderator. Pada threshold 0,35, model mencapai keseimbangan optimal antara sensitivitas deteksi dan ketepatan klasifikasi (precision 86,31%), menghasilkan F1-Score maksimal sebesar 0,8753.

Dalam konteks sistem keamanan digital, angka Recall ini sangat krusial karena merepresentasikan kemampuan model dalam melindungi korban dengan meminimalkan lolosnya konten berbahaya (FN). Di saat yang sama, model juga mempertahankan Precision sebesar 86,31%, yang berarti model cukup cerdas untuk tidak terlalu sering salah menuduh komentar netral sebagai bullying, menjaga kenyamanan pengguna dari moderasi yang berlebihan.Selain itu, ketangguhan model divalidasi oleh skor ROC-AUC (0,9291) dan PR-AUC (0,9425). Tingginya nilai PR-AUC secara khusus mengungguli dalam konteks dataset yang digunakan bahwa model mempertahankan performa klasifikasi yang stabil meskipun dihadapkan pada dataset yang tidak seimbang (imbalanced).

Dari total 1.455 data cyberbullying yang ada pada data uji, model berhasil mengklasifikasikan 1.292 data dengan benar (TP) dan hanya melewatkan 163 data (FN). Sementara itu, pada kelas non-cyberbullying, model berhasil memprediksi 959 data secara tepat. Untuk memberikan gambaran klasifikasi secara mendetail, distribusi prediksi model divisualisasikan melalui Confusion Matrix pada Gambar 4.



Gambar 5. *Visualisasi Confusion Matrix Model Usulan*

Kinerja tinggi ini dipengaruhi oleh integrasi pre-trained model IndoBERTweet yang telah memahami konteks bahasa gaul dan singkatan khas Twitter Indonesia, dikombinasikan dengan lapisan BiLSTM yang mampu menangkap ketergantungan urutan kata secara dua arah (maju dan mundur). Hal ini memungkinkan model untuk membedakan antara opini netral dan serangan personal yang bersifat halus.

3.4 Perbandingan dengan Model Baseline

Tahap terakhir dari analisis hasil adalah membandingkan performa model usulan (IndoBERTweet-BiLSTM) dengan dua model baseline lainnya, yaitu IndoBERT-BiLSTM dan IndoBERT-P1-BiLSTM. Perbandingan ini bertujuan untuk melihat efektivitas penggunaan pre-trained model yang dispesialisasi untuk data media sosial (IndoBERTweet) dibandingkan dengan model bahasa Indonesia umum. Hasil perbandingan secara menyeluruh disajikan pada Tabel 6.

Tabel 6. Perbandingan Kinerja Model Usulan vs Baseline (Kelas Cyberbullying)

Model	Akurasi	F1-Score	Recall	ROC-AUC	PR-AUC
IndoBERTweet-BiLSTM	0.8595	0.8753	0.8880	0.9291	0.9425
IndoBERT-BiLSTM	0.7934	0.8290	0.9010	0.8723	0.8820
IndoBERT-p1-BiLSTM	0.8381	0.8632	0.9196	0.9201	0.9317

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9249

Berdasarkan Tabel 6, model usulan IndoBERTweet-BiLSTM mengungguli kedua model lainnya pada hampir seluruh metrik utama, khususnya pada Akurasi (0,8595), F1-Score (0,8753), ROC-AUC (0,9291), dan PR-AUC (0,9425). Hal ini menunjukkan bahwa pengetahuan yang dimiliki IndoBERTweet mengenai karakteristik bahasa Twitter (seperti penggunaan slang, singkatan, dan emoji) sangat krusial dalam mendeteksi nuansa perundungan digital di Indonesia.

Keunggulan model usulan terlihat secara konsisten pada metrik ROC-AUC dan PR-AUC yang mencapai nilai tertinggi. Tingginya nilai PR-AUC (0,9425) menjadi indikator penting mengingat konteks dataset yang memiliki ketidakseimbangan kelas ringan. Metrik PR-AUC lebih sensitif terhadap performa model pada kelas minoritas dibandingkan ROC-AUC, sehingga nilai 0,9425 mengonfirmasi bahwa model usulan tidak hanya mampu membedakan kelas dengan baik (ROC-AUC tinggi), tetapi juga mempertahankan precision yang stabil pada berbagai level recall kemampuan yang esensial untuk sistem moderasi konten yang menuntut akurasi tinggi dengan minimalisasi false alarm.

Meskipun demikian, model IndoBERT-P1-BiLSTM menunjukkan performa yang sangat kompetitif dengan Recall tertinggi (0,9196) dan PR-AUC sebesar 0,9317. Hal ini mengindikasikan bahwa model tersebut sangat agresif dalam mengenali konten perundungan, namun memiliki presisi yang sedikit lebih rendah dibandingkan model usulan karena kecenderungan memberikan label positif yang lebih banyak. Di sisi lain, model IndoBERT-BiLSTM (versi standar) memberikan performa terendah dengan akurasi di bawah 80% dan PR-AUC sebesar 0,8820, yang memperkuat argumen bahwa model bahasa umum kurang optimal jika diterapkan langsung pada domain media sosial yang informal tanpa spesialisasi data yang mendalam.

Secara keseluruhan, integrasi arsitektur BiLSTM pada ketiga model terbukti mampu memberikan hasil yang stabil, namun pemilihan backbone IndoBERTweet tetap memberikan hasil yang paling seimbang dan akurat untuk kasus deteksi cyberbullying, sebagaimana ditunjukkan oleh superioritas konsisten pada seluruh metrik evaluasi termasuk PR-AUC yang menjadi ukuran paling reliabel untuk dataset tidak seimbang.

3.5 Analisis Kesalahan Klasifikasi

Untuk memahami batasan kinerja model IndoBERTweet-BiLSTM secara lebih mendalam, dilakukan analisis kualitatif terhadap sampel data yang mengalami kesalahan klasifikasi (misclassification) berdasarkan hasil pengujian. Dari total 2.619 data uji, model menghasilkan 368 kesalahan prediksi (14,05% dari total data uji) yang terdiri dari 205 kasus FP (7,83%) dan 163 kasus FN (6,22%). Analisis terhadap kedua jenis kesalahan ini memberikan wawasan penting mengenai tantangan linguistik yang masih dihadapi model dalam mendeteksi nuansa perundungan digital di media sosial Indonesia.

Kesalahan tipe FP terjadi ketika model memprediksi suatu konten sebagai Cyberbullying padahal sebenarnya Non-Cyberbullying. Dengan 205 kasus yang tercatat, jenis kesalahan ini menjadi dominan dibandingkan FN. Analisis kualitatif terhadap sampel-sampel dengan tingkat confidence tertinggi mengungkapkan beberapa pola kesalahan yang konsisten, sebagaimana ditunjukkan pada Tabel 7.

Tabel 7. Contoh Kasus FP dengan Confidence Tinggi

No	Teks Tweet (Original)	Confidence	Analisis Penyebab Kesalahan
1	"Serem cuy liat bencong berantem, jiwa lakinya keluar"	0.9879	Penggunaan kata "bencong" yang bermakna kasar, namun konteksnya bersifat observasional (mengomentari situasi), bukan serangan personal.
2	"Kapolri Tito: Polisi Jadi Bandar Narkoba, Tembak Mati Saja !!"	0.9765	Kalimat imperatif "tembak mati" terdeteksi sebagai ujaran agresif, padahal ini merupakan kutipan statement pejabat publik (berita).
3	"Intoleransi adalah mitos bagi komunis yg takut agama!! Teroris adalah mitos alat komunis..."	0.9750	Kata-kata seperti "komunis", "teroris", dan tanda seru ganda memicu klasifikasi agresif, padahal ini adalah opini politik tanpa serangan personal.

Berdasarkan Tabel 7, teridentifikasi tiga pola utama kesalahan FP. Pola pertama adalah Umpatan Non-Directed (Non-Targeted Profanity), di mana model menangkap keberadaan kata kasar seperti “bencong” atau frasa agresif, namun gagal membedakan apakah kata tersebut ditujukan untuk menyerang individu atau sekadar ekspresi emosi terhadap situasi. Pada Sampel #1, tweet bersifat naratif tentang pengamatan situasi, bukan serangan personal. Pola kedua berkaitan dengan Konteks Berita dan Kutipan, sebagaimana terlihat pada Sampel #2 yang menunjukkan bahwa model kesulitan membedakan antara kutipan pernyataan pejabat (yang dapat mengandung bahasa tegas) dengan ujaran kebencian aktual. Model menangkap kata “tembak mati” sebagai indikator agresivitas tanpa memahami bahwa ini adalah konteks pemberitaan. Pola ketiga adalah Opini Politik Keras, di mana Sampel #3 mengandung kritik keras terhadap ideologi politik dengan penggunaan terminologi sensitif “komunis”, “teroris”). Model cenderung mengklasifikasikan kritik politik yang menggunakan bahasa kuat sebagai cyberbullying, meskipun tidak ada serangan terhadap individu spesifik.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9249

Di sisi lain, kesalahan tipe FN terjadi ketika model memprediksi konten sebagai Non-Cyberbullying padahal sebenarnya mengandung unsur Cyberbullying. Dengan 163 kasus yang tercatat (6,22% dari data uji), jenis kesalahan ini lebih krusial dari sisi keamanan karena berkaitan dengan lolosnya konten berbahaya yang berpotensi merugikan korban. Analisis terhadap sampel dengan confidence terendah (paling yakin aman, namun ternyata berbahaya) mengungkapkan pola-pola kegagalan deteksi sebagaimana ditampilkan pada Tabel 8.

Tabel 8. Contoh Kasus FN dengan Confidence Rendah

No	Teks Tweet (Original)	Confidence	Analisis Penyebab Kesalahan
1	"Dg niat ridho & ikhlas, saya berjanji jika JOKOWI kalah di pilpres 2019 saya akan Sujud syukur di Tengah2 Alun-alun"	0.0058	Perundungan implisit melalui doa negatif. Tidak ada kata kasar eksplisit, namun mengandung hasutan dan kebencian terselubung terhadap tokoh politik.
2	"pemimpindarimasjid adalah pemimpin yg dapat mengajak kita menjalankan syariat Islam. Jatuhkan pemimpin non muslim"	0.0072	Hasutan dengan narasi religius yang mengandung diskriminasi dan provokasi politik, namun dikemas dalam bahasa formal tanpa profanitas.
3	"Saya kirain ada orang yg bisa ke bulan naik Onta. Setau saya USA & Yahudi kebanyakan pake Otak bukan pake Onta 😂"	0.0133	Sarkasme halus dengan stereotip etnis. Perbandingan implisit yang merendahkan ("Onta" vs "Otak") dan nada mengejek yang dikemas dalam struktur kalimat tidak agresif menyamarkan intensi diskriminatif terhadap kelompok tertentu.

Berdasarkan Tabel 8, teridentifikasi tiga pola utama kesalahan False Negative. Pola pertama adalah Perundungan Implisit Berbasis Ideologi, sebagaimana terlihat pada Sampel #1 dan #2 yang menunjukkan pola serangan yang tidak menggunakan kata kasar eksplisit, namun mengandung hasutan politik dan diskriminasi religius. Model yang dilatih untuk mendeteksi profanitas dan serangan langsung gagal mengenali ujaran kebencian yang dikemas dalam bahasa formal atau religius. Pola kedua berkaitan dengan Sarkasme dan Stereotip Etnis, di mana Sampel #3 menggunakan gaya bahasa sindiran dengan perbandingan stereotip etnis ("Onta" vs "Otak"). Struktur kalimat yang tidak mengandung kata kasar eksplisit dan nada sarkasme yang halus menipu model untuk mengklasifikasikannya sebagai komentar biasa, padahal mengandung intensi merendahkan kelompok etnis tertentu. Pola ketiga adalah Keterbatasan Deteksi Konteks Pragmatik, di mana model kesulitan memahami intensi dan implikatur di balik kalimat yang secara sintaksis dan leksikal terlihat netral, namun secara pragmatik mengandung hasutan atau diskriminasi.

Temuan ini mengindikasikan bahwa meskipun IndoBERTweet memiliki pemahaman yang baik terhadap bahasa informal Twitter, model masih memerlukan pemahaman pragmatik yang lebih dalam untuk membedakan antara ekspresi kasar yang bersifat ekspresif dengan serangan personal yang sesungguhnya. Analisis kesalahan ini memberikan beberapa implikasi penting untuk pengembangan model di masa mendatang. Pertama, terdapat kebutuhan mendesak untuk augmentasi data kontekstual, di mana dataset pelatihan perlu diperkaya dengan sampel-sampel yang mengandung sarkasme, sindiran halus, dan ujaran kebencian berbasis ideologi/religius yang tidak menggunakan profanitas eksplisit. Kedua, integrasi pemahaman pragmatik menjadi sangat krusial, sehingga arsitektur model dapat diperkuat dengan mekanisme attention yang lebih kompleks atau integrasi knowledge graph untuk memahami konteks sosial-politik dan intensi di balik teks. Ketiga, review kualitas anotasi dataset perlu dilakukan untuk memastikan konsistensi pelabelan, terutama pada kasus-kasus ambigu yang berada di area abu-abu antara ekspresi emosi dan serangan personal. Keempat, penyesuaian threshold berbasis konteks aplikasi menjadi pertimbangan penting, terutama untuk sistem moderasi konten yang memprioritaskan keamanan pengguna (menghindari lolosnya konten berbahaya), di mana threshold dapat diturunkan untuk meningkatkan Recall meskipun dengan trade-off peningkatan FP yang dapat direview secara manual.

Secara keseluruhan, meskipun model IndoBERTweet-BiLSTM menunjukkan performa yang sangat baik dengan tingkat akurasi 85,95%, analisis kesalahan mengungkapkan bahwa tantangan utama terletak pada deteksi nuansa bahasa yang sangat halus terutama perundungan implisit, sarkasme, dan ujaran kebencian yang dikemas dalam narasi ideologis formal.

4. KESIMPULAN

Penelitian ini berhasil mengembangkan sistem deteksi cyberbullying berbahasa Indonesia menggunakan pendekatan hibrida IndoBERTweet-BiLSTM yang diperkuat dengan strategi Class Weighting dan Label Smoothing, mencapai F1-Score 0,8753 dengan akurasi 85,95%, Precision 86,31%, Recall 88,80%, dan ROC-AUC 0,9291 pada dataset 13.091 tweet. Keunggulan model usulan terbukti secara komparatif melalui margin signifikan terhadap baseline IndoBERT-BiLSTM (79,34%) dan IndoBERT-P1-BiLSTM (83,81%), mengonfirmasi pentingnya domain alignment dalam memahami karakteristik bahasa informal media sosial. Meskipun demikian, analisis terhadap 368 kesalahan

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9249

klasifikasi mengungkapkan keterbatasan model dalam menangani nuansa halus, di mana 205 kasus False Positive (7,83%) terutama berasal dari umpatan non-directed, kutipan berita tegas, dan opini politik keras yang salah teridentifikasi sebagai cyberbullying, sementara 163 kasus False Negative (6,22%) menunjukkan kegagalan deteksi pada perundungan implisit berbasis ideologi, sarkasme stereotip etnis, dan ujaran kebencian yang memerlukan pemahaman pragmatik mendalam. Temuan ini mengindikasikan perlunya pengembangan lebih lanjut melalui augmentasi data kontekstual untuk kasus sarkasme dan sindiran halus, integrasi mekanisme attention multi-head atau knowledge graph untuk reasoning pragmatik, review kualitas anotasi pada kasus ambigu, serta eksplorasi penyesuaian threshold berbasis konteks aplikasi untuk optimasi trade-off Recall-Precision sesuai prioritas moderasi. Secara keseluruhan, penelitian ini memberikan kontribusi signifikan dalam pengembangan sistem deteksi cyberbullying berbahasa Indonesia yang akurat dan efisien, sekaligus membuka peluang riset lanjutan untuk menangani kompleksitas linguistik subtil di media sosial.

REFERENCES

[1] P. Yi and A. Zubiaga, "Session-based cyberbullying detection in social media: A survey," *Online Soc Netw Media*, vol. 36, Jul. 2023, doi: 10.1016/j.osnem.2023.100250.

[2] A. Perera and P. Fernando, "Accurate cyberbullying detection and prevention on social media," *Procedia Comput Sci*, vol. 181, pp. 605–611, 2021, doi: 10.1016/j.procs.2021.01.207.

[3] S. Singh and S. H. Othman, "An Effective Cyberbullying Detection Model for the Malay Language Using Transformer Model in Social Media Platform X," *International Journal of Innovative Computing*, vol. 15, no. 1, pp. 63–71, May 2025, doi: 10.11113/ijic.v15n1.520.

[4] S. Sihab-Us-Sakib, Md. R. Rahman, Md. S. A. Forhad, and Md. A. Aziz, "Cyberbullying detection of resource constrained language from social media using transformer-based approach," *Natural Language Processing Journal*, vol. 9, p. 100104, Dec. 2024, doi: 10.1016/j.nlp.2024.100104.

[5] J. Forry Kusuma and A. Chowanda, "Indonesian Hate Speech Detection Using IndoBERTweet and BiLSTM on Twitter," *JOIV: International Journal on Informatics Visualization*, vol. 7, no. 3, pp. 773–780, 2023, doi: 10.30630/joiv.7.3.1035.

[6] P. H. Zakaria, D. Nurjannah, and H. Nurrahmi, "Misogyny Text Detection on Tiktok Social Media in Indonesian Using the Pre-trained Language Model IndoBERTweet," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 7, no. 3, pp. 1297–1305, Jul. 2023, doi: 10.30865/mib.v7i3.6438.

[7] G. Z. Nabiilah, S. Y. Prasetyo, Z. N. Izdihar, and A. S. Girsang, "BERT base model for toxic comment analysis on Indonesian social media," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 714–721. doi: 10.1016/j.procs.2022.12.188.

[8] K. H. Yunior, A. V. Vitianingsih, S. Kacung, A. Lidya Maukar, and A. Dwi Arumsari, "Sentiment Analysis of Cyberbullying Detection on Social Networks using the Sentistrenght Method," *Sistemasi: Jurnal Sistem Informasi*, vol. 13, no. 4, pp. 1587–1596, 2024, doi: 10.32520/stmsi.v13i4.4226.

[9] S. A. B. Sibarani, R. Purba, and R. P. Limbong, "Implementation of IndoBERT in Sarcasm Detection using Random Forest Towards Sentiment Analysis," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 4, pp. 2120–2130, Mar. 2025, doi: 10.47065/bits.v6i4.5801.

[10] A. Baruah, K. A. Das, F. A. Barbhuiya, and K. Dey, "Context-Aware Sarcasm Detection Using BERT," in *Proceedings of the Second Workshop on Figurative Language Processing*, Association for Computational Linguistics, Jul. 2020, pp. 24–29. doi: 10.18653/v1/P17.

[11] E. Scola and I. Segura-Bedmar, "Sarcasm Detection with BERT," *Procesamiento del Lenguaje Natural*, vol. 67, pp. 13–25, Sep. 2021, doi: 10.26342/2021-67-1.

[12] T. Javed, M. A. Nouman, and R. Zahid, "BERT Model Adoption for Sarcasm Detection on Twitter Data," *VFAST Transactions on Software Engineering*, vol. 12, no. 3, pp. 177–198, Sep. 2024, doi: 10.21015/vtse.v12i3.1908.

[13] A. Kannammal, S. Omprakash, and J. D. Dheerthan, "Automated Decision Support System for Cyberbullying Detection," in *Procedia Computer Science*, Elsevier B.V., 2023, pp. 760–768. doi: 10.1016/j.procs.2023.12.130.

[14] F. R. Sayed, E. H. Elnashar, and F. A. Omara, "Cyberbullying Detection in Social Media Using Natural Language Processing," *Sci Afr*, p. e02713, Apr. 2025, doi: 10.1016/j.sciaf.2025.e02713.

[15] A. G, H. Kumar, and B. D, "Toxic Comment Classification using Transformers," in *Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management*, Singapore, 2021, pp. 1895–1905. doi: 10.46254/AN11.20210366.

[16] F. Shely Amalia and Y. Suyanto, "Offensive language and hate speech detection using BERT model," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 18, no. 1, 2024, doi: 10.22146/ijccs.99841.

[17] F. Rahman and A. S. Girsang, "IndoBERTweet for Sarcasm: Evaluating Domain-Adapted Transformers for Indonesian Twitter Sarcasm Classification," *Journal of Logistics, Informatics and Service Science*, vol. 11, no. 2, pp. 155–164, 2024, doi: 10.33168/JLISS.2024.0210.

[18] S. Kaya and B. Alatas, "Sarcasm Detection with A New CNN+BiLSTM Hybrid Neural Network and BERT Classification Model," *International Journal of Advanced Networking and Applications*, vol. 14, no. 03, pp. 5436–5443, 2022, doi: 10.35444/ijana.2022.14304.

[19] A. Amudhan, A. G, R. S, and S. Niveditha, "Toxic comment classification," *International Research Journal of Modernization in Engineering Technology and Science (IRJMETS)*, vol. 6, no. 10, pp. 2093–2099, 2024, doi: 10.56726/IRJMETS62348.

[20] Y. M. Ibrahim, R. Essameldin, and S. M. Saad, "Social Media Forensics: An Adaptive Cyberbullying-Related Hate Speech Detection Approach Based on Neural Networks With Uncertainty," *IEEE Access*, vol. 12, pp. 59474–59484, 2024, doi: 10.1109/ACCESS.2024.3393295.

[21] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," in *Proceedings of the 3rd Workshop on Abusive Language Online (ALW3)*, Florence, Italy: Association for Computational Linguistics, 2019, pp. 46–57. doi: 10.18653/v1/W19-3506.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9249

[22] J. Khan, K. Ahmad, S. K. Jagatheesaperumal, and K. A. Sohn, "Textual variations in social media text processing applications: challenges, solutions, and trends," *Artif Intell Rev*, vol. 58, no. 89, Mar. 2025, doi: 10.1007/s10462-024-11071-z.

[23] X. Xie, M. Xie, A. J. Moshayedi, and M. H. Noori Skandari, "A Hybrid Improved Neural Networks Algorithm Based on L2 and Dropout Regularization," *Math Probl Eng*, vol. 2022, pp. 1–19, 2022, doi: 10.1155/2022/8220453.

[24] A. A. Hafiza and E. B. Setiawan, "Enhancing Cyberbullying Detection on Platform 'X' Using IndoBERT and Hybrid CNN-LSTM Model," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 2, pp. 655–672, Apr. 2025, doi: 10.52436/1.jutif.2025.6.2.4321.

[25] Moh. H. Fariz and E. B. Setiawan, "The Impact of Word Embedding on Cyberbullying Detection Using Hybrid Deep Learning CNN-BiLSTM," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 10, no. 3, pp. 661–671, Feb. 2025, doi: 10.33480/jitk.v10i3.6270.