Template-Jurnal-UMSIDA-new (3)

11% Suspicious texts

2% Similarities
2 % similarities between quotation marks
0 % among the sources mentioned
9% Unrecognized languages
34% Texts potentially generated by AI (ignored)

Document name: Template-Jurnal-UMSIDA-new (3).docx
Document ID: c8630ec944bbeb38888281e6101c46949549a425
Original document size: 499.07 KB

Submitter: fst umsida
Submission date: 1/30/2026
Upload type: interface
analysis end date: 1/30/2026

Number of words: 9,321
Number of characters: 74,072

Location of similarities in the document:

Sources of similarities

Main sources detected

No.	Description	Similarities	Locations	Additional information
1	dx.doi.org | VISUALISASI ANALISIS SENTIMEN SIBERBULLYING PADA POST INSTAG... http://dx.doi.org/10.59095/ijcsr.v2i1.15	< 1%		Identical words: < 1% (30 words)
2	sistemasi.ftik.unisi.ac.id | Applying Artificial Intelligence to Analyze Emotions in ... https://sistemasi.ftik.unisi.ac.id/index.php/stmsi/article/download/5187/981	< 1%		Identical words: < 1% (26 words)
3	dx.doi.org | Analisis Sentimen Review Hotel Menggunakan Metode Deep Learnin... http://dx.doi.org/10.24002/jbi.v14i02.7244	< 1%		Identical words: < 1% (23 words)

Sources with incidental similarities

No.	Description	Similarities	Locations	Additional information
1	dx.doi.org | Deteksi Komentar Cyberbullying Pada YouTube Dengan Metode Con... http://dx.doi.org/10.34148/teknika.v12i3.677	< 1%		Identical words: < 1% (27 words)
2	journal.trunojoyo.ac.id https://journal.trunojoyo.ac.id/edutic/article/download/28822/10616	< 1%		Identical words: < 1% (14 words)
3	journal.goresearch.id http://journal.goresearch.id/index.php/sij/article/view/197	< 1%		Identical words: < 1% (19 words)
4	doi.org | Implementasi Data Mining Untuk Klasifikasi Komentar Hate Speech Men... https://doi.org/10.30591/smartcomp.v14i3.8122	< 1%		Identical words: < 1% (19 words)
5	dx.doi.org | Perbandingan analisis sentimen pada aplikasi SIREKAP dengan aplika... http://dx.doi.org/10.33022/ijcs.v13i4.4084	< 1%		Identical words: < 1% (16 words)

Referenced sources (without similarities detected)

These sources were cited in the paper without finding any similarities.

1 https://t.co/NLTPgeBuka
2 https://t.co/ewhKdZIGGa
3 https://dataindonesia.id/internet/detail/pengguna-media-sosial-diindonesia-sebanyak-167-juta-pada-2023
4 https://jurnal.umj.ac.id/index.php/khidmatsosial/article/view/10433
5 https://www.taylorfrancis.com/chapters/edit/10.4324/9780203848166-22/mothers-fathers-coparenting-together-john-beaton-william-doherty-lisa-wenger

Points of interest

INDOBERTWEET-BILSTM FOR DETECTING CYBERBULLYING IN MIXED INDONESIAN-ENGLISH SOCIAL MEDIA TEXTS
[INDOBERTWEET-BILSTM UNTUK DETEKSI CYBERBULLYING PADA TEKS MEDIA SOSIAL CAMPURAN BAHASA INDONESIA-INGGRIS]

Mochammad Wahyu Alvy Kusuma1), Mochamad Alfan Rosid*1), Sukma Aji1), Suhendro Busono1)

1) Departemen Informatika, Fakultas Sains dan Teknologi, Universitas Muhammadiyah Sidoarjo, Indonesia
*Email Penulis Korespondensi: alfanrosid@umsida.ac.id

Abstract.

Cyberbullying has become a serious issue on Indonesian social media, with approximately 45% of teenagers having experienced it. Platform X exhibits a high rate of cyberbullying, complicated by the code-mixing phenomenon between Indonesian and English. This study develops an optimal cyberbullying detection system through systematic hyperparameter configuration investigation and comprehensive deep learning architecture comparison. We evaluated three optimizers (Adam, AdamW, SGD) with two learning rate schedulers (Linear and Cosine Annealing), and compared six deep learning architectures with one traditional machine learning method. The dataset consisted of 13,677 code-mixed Indonesian tweets (80% training, 10% validation, 10% testing). The IndoBERTweet + BiLSTM model with AdamW optimizer and Cosine Annealing scheduler achieved the best performance: 91.48% accuracy, precision, recall, and F1-score. Adaptive optimizers showed significant impact (19.22% gap vs. SGD), while learning rate schedulers provided consistent improvement of 0.33%. IndoBERTweet improved accuracy by 2.7%-3.8%, outperforming the SVM + TF-IDF baseline (78.18%) by 13.3%.
Keywords - BiLSTM;
Code-mixed; Cyberbullying detection; Deep learning; Hyperparameter optimization;
IndoBERTweet; Sentiment analysis

Abstrak. Perundungan siber (cyberbullying) telah menjadi isu serius di media sosial Indonesia, dengan sekitar 45% remaja pernah mengalaminya. Platform X memiliki tingkat cyberbullying tinggi, diperumit oleh fenomena code-mixing bahasa Indonesia-Inggris. Penelitian ini mengembangkan sistem deteksi cyberbullying optimal melalui investigasi konfigurasi hyperparameter dan perbandingan arsitektur deep learning. Kami mengevaluasi tiga optimizer (Adam,
AdamW, SGD) dengan dua learning rate scheduler (Linear dan Cosine Annealing),
serta membandingkan enam arsitektur deep learning dan satu metode machine learning tradisional. Dataset terdiri dari 13.677 tweet code-mixed berbahasa Indonesia (80% latih, 10% validasi, 10% uji). Model IndoBERTweet + BiLSTM dengan optimizer AdamW dan scheduler Cosine Annealing mencapai performa terbaik: akurasi 91,48%, precision 91,48%, recall 91,48%, dan F1-score 91,48%.

Adaptive optimizer memberikan pengaruh signifikan (gap 19,

22% vs SGD), sedangkan learning rate scheduler meningkat 0,33%. IndoBERTweet meningkatkan akurasi 2,

7%-3,8%, unggul 13,3% dibanding baseline SVM + TF-IDF (78,18%).

Kata Kunci - Analisis sentimen; BiLSTM; Code-mixed; Deep learning; Deteksi cyberbullying; IndoBERTweet; optimasi hiperparameter;

I. Pendahuluan

Penggunaan media sosial telah menjadi bagian penting dalam kehidupan masyarakat modern. Platform ini tidak hanya dimanfaatkan untuk berkomunikasi antar individu, tetapi juga untuk berbagai keperluan lain seperti pendidikan, hiburan, dan bisnis. Media sosial didefinisikan sebagai platform berbasis internet yang memungkinkan pengguna berinteraksi, memberikan komentar, dan merespons unggahan pengguna lain secara langsung [1].
Media sosial X (sebelumnya Twitter), yang memungkinkan pengguna berbagi informasi dan ber-interaksi secara waktu nyata, merupakan salah satu platform yang sangat populer. Berdasarkan data We Are Social tahun 2023, jumlah pengguna media sosial aktif di Indonesia mencapai 167 juta orang [2], dengan media sosial X menjadi salah satu platform yang banyak digunakan. Fitur anonimitas dan kemu-dahan akses pada platform ini berpotensi memicu dampak negatif, salah satunya meningkatnya kasus cyberbullying atau perundungan siber.
Cyberbullying merupakan bentuk perundungan yang dilakukan berulang melalui media elektronik terhadap seseorang yang tidak mudah membela diri [3].

Fenomena ini menunjukkan tren peningkatan di Indonesia, dengan laporan UNICEF tahun 2023 yang menyatakan sekitar 45% remaja Indonesia pernah mengalami cyberbullying di media sosial [4]. Dampaknya dapat sangat serius, mulai dari gangguan psikologis, depresi, hingga kasus bunuh diri yang dilaporkan berbagai media [5][6].
Seiring meningkatnya kasus cyberbullying, dibutuhkan sistem yang mampu menganalisis sentimen komentar di media sosial X secara otomatis dan efektif. Deteksi cyberbullying menggunakan analisis sen-timen telah menjadi topik penelitian yang berkembang pesat dalam beberapa tahun terakhir. Berbagai pen-dekatan telah dikembangkan, mulai dari metode Machine learning tradisional hingga arsitektur Deep learning yang lebih kompleks. Metode tradisional seperti Support Vector Machine (SVM) dengan TF-IDF [7],
Naive Bayes [8], dan Random Forest [9] telah banyak digunakan,
namun memiliki keterbatasan dalam memahami konteks dan makna semantik dari teks.

Perkembangan teknologi Deep learning, seperti Long Short-Term Memory (LSTM) [10] dan Convolutional Neural Network (CNN) [11],

menunjukkan pening-katan performa dalam menangkap pola sekuensial dan fitur kontekstual pada teks.

Penelitian awal dalam deteksi cyberbullying banyak menggunakan algoritma Machine learning klasik dengan representasi fitur berbasis TF-IDF. Abdullah dan Hidayatullah [12] membandingkan empat algoritma klasifikasi yaitu KNN, SVM, Logistic Regression, dan Naïve Bayes pada dataset Twitter dengan ekstraksi fitur menggunakan TF-IDF, di mana SVM menunjukkan performa terbaik dengan akurasi 99,75%.

Rizki, Auliasari, dan Prasetya [13] menerapkan SVM untuk analisis sentimen cyberbullying pada Twitter dan memperoleh akurasi 70%, menunjukkan bahwa meskipun SVM efektif untuk klasifikasi teks, masih terdapat ruang untuk peningkatan performa. Hasan [14] mengeksplorasi algoritma K-Nearest Neigh-bor (KNN) untuk deteksi cyberbullying di Facebook, dengan model 1-NN mencapai akurasi tertinggi 71,43%. Pendekatan ensemble learning juga telah diterapkan, seperti penelitian Nurnaryo, Mualaab, dsb.

[15] yang mengintegrasikan seleksi fitur Information Gain dengan Random Forest, dan Santoso, Putri, dan Sahbandi [16] yang menggunakan Random Forest dengan Word embedding FastText pada komentar Insta-gram berbahasa Indonesia, mencapai akurasi 84%. Penelitian dengan algoritma Naïve Bayes menunjukkan hasil yang bervariasi, di mana Baehaqi dan Cahyono [17] memperoleh akurasi 91,25% pada komentar In-stagram, sedangkan Machmud, Wibisono, dan Suryani [18] mencapai akurasi 80,77% pada dataset Twitter. Triyana, Putra, dan Pradhana [19] mengevaluasi SVM pada dataset tweet berbahasa Inggris yang terdiri dari 24.783 sampel, mencapai akurasi 90,59%, presisi 95,42%, dan Recall 92,74%, namun menyarankan eksplorasi lebih lanjut menggunakan model Deep learning untuk meningkatkan pemahaman konteks se-mantik yang menjadi keterbatasan metode tradisional. Perkembangan teknologi Deep learning membawa pendekatan baru dalam deteksi cyberbullying dengan kemampuan yang lebih baik dalam memahami konteks dan melakukan ekstraksi fitur secara otomatis. Widiyantoro, dan Prasetyo [20] menerapkan Long Short-Term Memory (LSTM) untuk mendeteksi cyberbullying terhadap pemain sepak bola di Twitter, mencapai akurasi awal 59,7% dengan F1-Score 45% yang kemudian ditingkatkan melalui Hyperparameter tuning. Andika, Kristian, Dan Se-tiawan [21] mengembangkan arsitektur hybrid CNN-LSTM untuk klasifikasi komentar cyberbullying di YouTube, memanfaatkan CNN untuk ekstraksi fitur spasial dan LSTM untuk menangkap dependensi sek-uensial dalam teks, menghasilkan F1-Score 0,84. Arsitektur yang lebih kompleks diusulkan oleh Rosid, Siahaan, dan Saikhu [22] yang menggabungkan CNN, multihead attention, dan Bidirectional GRU (BiGRU) untuk deteksi sarkasme dalam teks Code-mixed Indonesia-Inggris, mencapai akurasi 94,60% dengan F1-Score 94,38% pada dataset pertama, namun mengalami penurunan performa menjadi 88,02% pada dataset kedua, mengindikasikan bahwa generalisasi model pada dataset yang berbeda masih menjadi tantangan.

Model berbasis transformer, khususnya BERT (Bidirectional Encoder Representations from Trans-formers) [23],

telah membawa terobosan signifikan dalam pemahaman konteks bahasa melalui mekanisme attention bidireksional. Safitri, dsb. [24] membandingkan BERT dengan Random Forest untuk deteksi cyberbullying pada 650 tweet, dengan BERT mencapai akurasi 94%, melampaui metode Machine learning tradisional secara signifikan. Untuk konteks bahasa Indonesia di media sosial, model IndoBERTweet yaitu pengembangan dari BERT yang dioptimalkan untuk analisis teks berbahasa Indonesia di platform media sosial [25] menjadi salah satu pendekatan yang menjanjikan. Model ini dilatih secara khusus menggunakan data tweet berbahasa Indonesia, sehingga mampu menangkap karakteristik bahasa gaul, singkatan, dan ek-spresi informal yang sering muncul di media sosial X. Zakaria, Nurjannah, dan Nurrahmi [25] menggunakan IndoBERTweet untuk mendeteksi teks misogini pada komentar TikTok berbahasa Indonesia yang terdiri dari 1.576 komentar dan mencapai akurasi 76,89%. Kusuma, dan Chawonda [26] mengom-binasikan IndoBERTweet dengan BiLSTM untuk deteksi ujaran kebencian di Twitter, memperoleh akurasi 93,7% pada satu dataset dan 88,6% pada dataset lainnya, menunjukkan bahwa integrasi model Pre-trained dengan arsitektur recurrent neural network dapat meningkatkan performa deteksi, terutama dalam me-nangkap karakteristik bahasa informal dan singkatan yang sering muncul di media sosial. Namun, sebagian besar model yang ada masih terbatas pada teks berbahasa Indonesia yang relatif baku, dan belum secara eksplisit menangani fenomena Code-mixing (campuran bahasa) yang banyak ditemukan di platform X. Di Indonesia, banyak pengguna media sosial terutama kalangan muda yang menggunakan campuran bahasa Indonesia dan Inggris (Code-mixed text) untuk mengekspresikan diri secara lebih bebas dan kreatif. Selain itu, minimnya penelitian yang secara spesifik menangani teks Code-mixed Indonesia-Inggris untuk deteksi cyberbullying, padahal fenomena ini sangat prevalent di kalangan pengguna media sosial Indonesia terutama generasi muda. Beberapa penelitian mencapai akurasi tinggi namun tidak melaporkan secara detail proses optimasi Hyperparameter yang dilakukan, sehingga sulit un-tuk mereplikasi hasil atau memahami faktor-faktor yang berkontribusi terhadap performa model. Sebagian penelitian juga tidak membandingkan model yang diusulkan dengan berbagai baseline yang memadai, baik dari sisi Machine learning tradisional maupun arsitektur Deep learning alternatif.
Kesenjangan penelitian (research gap) dalam studi ini terletak pada terbatasnya sistem deteksi cyberbullying yang mampu memahami dan mengklasifikasikan sentimen dalam dataset yang terdiri atas teks Code-mixed Indonesia–Inggris. Model yang tidak dirancang untuk konteks ini cenderung kesulitan memproses kode campuran yang mengandung elemen linguistik yang berbeda dari bahasa baku. Penelitian ini bertujuan untuk mengisi gap tersebut dengan mengembangkan sistem deteksi cyberbullying berbasis IndoBERTweet yang dioptimalkan untuk menangani teks Code-mixed Indonesia-Inggris, disertai dengan evaluasi sistematis terhadap berbagai konfigurasi Hyperparameter dan perbandingan komprehensif dengan multiple baseline methods.

Secara khusus, penelitian ini bertujuan untuk: (1) mengembangkan dan men-gevaluasi sistem deteksi cyberbullying pada media sosial X menggunakan IndoBERTweet melalui analisis sentimen, (2) menyelidiki pengaruh konfigurasi Hyperparameter khususnya Optimizer (Adam, AdamW, dan SGD) dengan Learning rate scheduler (Linear dan Cosine Annealing) terhadap performa model, dan (3) membandingkan berbagai arsitektur Deep learning berbasis IndoBERTweet (IndoBERTweet + BiLSTM,
IndoBERTweet + BiGRU, IndoBERTweet standalone) dengan model Deep learning lain (CNN-LSTM, BiGRU standalone, BiLSTM standalone) dan metode Machine learning tradisional (SVM + TF-IDF) sebagai baseline.

Kontribusi utama penelitian ini mencakup empat aspek penting. Pertama, penelitian ini menghadir-kan dataset baru dengan karakteristik Code-mixed Indonesia-Inggris yang dilabeli secara manual untuk tu-gas deteksi cyberbullying. Kedua, penelitian ini melakukan investigasi sistematis terhadap pengaruh kon-figurasi Hyperparameter, khususnya kombinasi Optimizer dan Learning rate scheduler, yang memberikan insight mendalam tentang faktor-faktor krusial dalam optimasi model transformer untuk bahasa Indonesia informal.

Ketiga, penelitian ini menyajikan perbandingan komprehensif berbagai arsitektur Deep learning dan traditional Machine learning methods dalam konteks deteksi cyberbullying teks Code-mixed, mem-berikan panduan empiris untuk pemilihan arsitektur optimal. Keempat, penelitian ini mengembangkan model deteksi cyberbullying dengan

mengombinasikan IndoBERTweet dan BiLSTM, mengungguli metode-metode existing secara signifikan.
Struktur artikel ini disusun sebagai berikut. Bagian II menjelaskan metodologi penelitian secara rinci, mencakup pengumpulan data, Preprocessing, pembagian dataset, arsitektur model yang diusulkan, skenario pengujian, dan metrik evaluasi. Bagian III menyajikan hasil eksperimen dan pembahasan men-dalam, termasuk analisis performa berbagai konfigurasi Hyperparameter dan perbandingan arsitektur model.

II. Metode

Bagian ini menguraikan metodologi penelitian secara rinci dan sistematis, mencakup tahapan pengumpulan dan persiapan data, perancangan arsitektur model, konfigurasi eksperimen, hingga metrik evaluasi yang digunakan. Alur penelitian secara keseluruhan diilustrasikan pada Gambar 1.

Gambar 1. Tahapan Penelitian

Gambar 1. Tahapan Penelitian

A. Tahapan Penelitian

Penelitian ini dilakukan melalui serangkaian tahapan yang sistematis dan terstruktur untuk memastikan validitas dan reliabilitas hasil. Tahapan dimulai dari studi literatur(Previous Research) untuk memahami landasan teori dan penelitian terkait, dilanjutkan dengan pengumpulan data mentah(Data Crawling) dari media sosial X. Data tersebut kemudian melalui tahap pra-pemrosesan (Text Preprocessing) untuk membersihkan dan mengubahnya ke dalam format yang sesuai untuk training model. Selanjutnya, data dibagi menjadi set data latih, validasi, dan uji(Data Splitting) dengan proporsi yang telah ditentukan. Model IndoBERTweet kemudian di-Fine-tune menggunakan data latih dan divalidasi secara berkala menggunakan data validasi untuk monitoring Convergence dan mencegah Overfitting.

Tahap akhir adalah pengujian dan evaluasi model menggunakan data uji yang Completely unseen untuk mengukur performa generalisasi dan menarik kesimpulan.
Studi literatur dilakukan secara komprehensif untuk meninjau dan menganalisis penelitian-penelitian sebelumnya yang relevan dengan deteksi cyberbullying dan analisis sentimen. Fokus utama tinjauan pustaka adalah pada topik analisis sentimen untuk teks berbahasa Indonesia, khususnya yang berasal dari media sosial X, serta penerapan berbagai arsitektur Natural Language Processing (NLP) seperti Transformer dan model BERT beserta variannya. Tujuan dari tinjauan pustaka ini adalah untuk membangun landasan teori yang kuat, mengidentifikasi research gap yang ada dalam literatur existing, dan memvalidasi pemilihan metode yang akan digunakan. Hasil dari tahap ini mengarahkan pada keputusan untuk mengadopsi model IndoBERTweet, yang telah terbukti memiliki performa unggul untuk pemrosesan teks dari Twitter berbahasa Indonesia karena Pre-training-nya yang spesifik pada domain Twitter Indonesia.

B. Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data teks berupa cuitan (tweet) berbahasa Indonesia yang diambil dari platform media sosial X (sebelumnya Twitter). Proses pengumpulan data dilakukan menggunakan Twitter API dengan menargetkan cuitan yang relevan dengan topik penelitian melalui kata kunci (keyword) atau tagar (hashtag) tertentu yang berkaitan dengan cyberbullying dan interaksi negatif di media sosial. Kriteria data yang dikumpulkan meliputi beberapa aspek penting: periode waktu pengambilan data ditentukan untuk memastikan relevansi temporal, dan memastikan cuitan yang diambil adalah cuitan asli (original tweets, bukan retweet) untuk menjaga orisinalitas dan kualitas data. Selain itu, tweet yang dikumpulkan dipastikan memiliki panjang yang memadai dan tidak hanya berisi mention atau link tanpa konten substantif. Contoh hasil dataset mentah dapat dilihat pada Tabel 1.

 Tabel 1. Contoh Hasil Data Mentah
No User Text
1 User1 Goblok ini Hoaxs ..Tolong Kalo Jadi Pendukung Anies terus Jangan Gilak Bikan Hoaxs .. Tolong @DivHumas_Polri ini ditindaklanjuti .
2 User2 Dialog Lingkungan Hidup Ustadz Abdul Somad dan Rocky Gerung di Tanjung Belit https://t.co/NLTPgeBuka
3 User3 Gibran Datang ke Rumahnya Rocky Gerung: Saya Kasih Kopi Oke You Bicara Anak Muda! https://t.co/ewhKdZIGGa

Tabel 1. Contoh Hasil Data Mentah
No User Text
1 User1 Goblok ini Hoaxs ..Tolong Kalo Jadi Pendukung Anies terus Jangan Gilak Bikan Hoaxs .. Tolong @DivHumas_Polri ini ditindaklanjuti .
2 User2 Dialog Lingkungan Hidup Ustadz Abdul Somad dan Rocky Gerung di Tanjung Belit https://t.co/NLTPgeBuka
3 User3 Gibran Datang ke Rumahnya Rocky Gerung: Saya Kasih Kopi Oke You Bicara Anak Muda! https://t.co/ewhKdZIGGa

Dataset akhir terdiri dari 13.677 tweet yang mencerminkan karakteristik komunikasi autentik di media sosial Indonesia. Data mentah yang terkumpul kemudian dilabeli secara manual melalui proses anotasi yang Rigorous. Proses Labeling ini sangat penting karena penelitian ini menggunakan pendekatan supervised learning (pembelajaran terarah) yang membutuhkan data berlabel sebagai Ground truth bagi model. Setiap cuitan dibaca dengan teliti dan diberi label sesuai dengan kategori yang telah ditentukan: label 1 untuk cyberbullying dan label 0 untuk non-cyberbullying.

Untuk menjaga konsistensi dan reliabilitas Labeling, sebuah pedoman anotasi (annotation guideline) disusun secara komprehensif sebagai acuan selama proses ini. Pedoman anotasi mencakup definisi operasional cyberbullying, contoh-contoh kasus, dan decision rules untuk kasus-kasus Ambiguous. Proses anotasi dilakukan oleh [jumlah anotator] yang terlatih, dengan mekanisme inter-rater agreement untuk memastikan konsistensi. Contoh dataset yang sudah dilabeli dapat dilihat pada Tabel 2.

 Tabel 2. Contoh Hasil Data Labeling
No Text Label
1 Goblok ini Hoaxs ..Tolong Kalo Jadi Pendukung Anies terus Jangan Gilak Bikan Hoaxs .. Tolong @DivHumas_Polri ini ditindaklanjuti . 1
2 Dialog Lingkungan Hidup Ustadz Abdul Somad dan Rocky Gerung di Tanjung Belit https://t.co/NLTPgeBuka 0
3 Gibran Datang ke Rumahnya Rocky Gerung: Saya Kasih Kopi Oke You Bicara Anak Muda! https://t.co/ewhKdZIGGa 0

Tabel 2. Contoh Hasil Data Labeling
No Text Label

1 Goblok ini Hoaxs ..Tolong Kalo Jadi Pendukung Anies terus Jangan Gilak Bikan Hoaxs .. Tolong @DivHumas_Polri ini ditindaklanjuti . 1
2 Dialog Lingkungan Hidup Ustadz Abdul Somad dan Rocky Gerung di Tanjung Belit https://t.co/NLTPgeBuka 0
3 Gibran Datang ke Rumahnya Rocky Gerung: Saya Kasih Kopi Oke You Bicara Anak Muda! https://t.co/ewhKdZIGGa 0

Karakteristik unik dari dataset ini adalah prevalensi tinggi penggunaan Code-mixing Indonesia-Inggris, di mana pengguna secara natural mencampurkan kedua bahasa dalam satu tweet. Selain itu, dataset juga mengandung bahasa gaul, slang, singkatan kreatif, typo yang disengaja, emoticon, dan berbagai variasi linguistik lainnya yang mencerminkan pola komunikasi informal di media sosial Indonesia. Distribusi kelas dalam dataset relatif seimbang untuk meminimalkan class imbalance issues yang dapat mempengaruhi performa model.

C. Preprocessing Data
Data yang telah dikumpulkan dan dilabeli selanjutnya melalui tahap pra-pemrosesan (Preprocessing). Tujuan utama dari tahap ini adalah untuk membersihkan data teks dari elemen-elemen yang tidak relevan (noise) yang dapat mengganggu proses klasifikasi, sehingga data menjadi siap untuk diolah oleh model Machine learning. Dalam penelitian ini, proses Preprocessing yang dilakukan mengadopsi pendekatan minimal Preprocessing yang berfokus pada data cleaning Essential.
Proses cleaning mencakup penghapusan Uniform Resource Locator (URL) yang sering muncul dalam tweet namun tidak memberikan informasi tekstual yang berguna untuk klasifikasi, dan penghapusan seluruh tanda baca yang terdapat dalam setiap data teks untuk mengurangi noise. Pendekatan minimal Preprocessing ini dipilih dengan pertimbangan khusus bahwa langkah pemrosesan yang lebih mendalam seperti stemming (pencarian kata dasar) atau Stopword removal (penghapusan kata umum) sengaja tidak diterapkan dalam penelitian ini.

Keputusan untuk tidak melakukan stemming dan Stopword removal didasarkan pada beberapa pertimbangan teoretis dan empiris yang penting. Pertama, model bahasa modern seperti yang digunakan dalam arsitektur IndoBERTweet seringkali bekerja lebih optimal dengan konteks kalimat yang lebih lengkap dan natural. Kedua, menghilangkan komponen teks secara berlebihan berisiko mengubah makna asli dari kalimat informal seperti tweet, di mana nuansa bahasa dan context words sangat penting untuk memahami intent dan sentiment. Ketiga, penelitian terdahulu pada Pre-trained language models menunjukkan bahwa models seperti BERT dan variannya memiliki kemampuan inheren untuk menangani stopwords dan berbagai inflections melalui contextual embeddings mereka, sehingga Preprocessing agresif justru dapat menurunkan performa dan akurasi model.
Keempat, dalam konteks cyberbullying detection,
kata-kata yang biasanya dianggap sebagai stopwords (seperti "lo", "gue", "kan") atau variasi kata (seperti "bego", "bodoh", "kebodohan") seringkali mengandung informasi penting tentang tone dan sentiment.

Setelah melalui proses cleaning yang minimal namun targeted ini, dataset yang bersih siap untuk dilanjutkan ke tahap pembagian data.

D. Pembagian Data
Dataset yang telah melalui tahap pra-pemrosesan kemudian dibagi menjadi tiga bagian (subset) yang berbeda dengan tujuan yang spesifik untuk memastikan evaluasi model yang fair dan reliable. Pembagian dilakukan menggunakan stratified random sampling untuk memastikan bahwa distribusi kelas (cyberbullying vs non-cyberbullying) konsisten di setiap subset, sehingga tidak ada bias dalam representasi kelas pada training, validation, atau test set.

Bagian terbesar adalah Data Latih (Training Data), yang mencakup 80% dari total dataset atau sekitar 8.752 tweet. Data latih digunakan secara langsung untuk melatih model agar dapat mengenali pola-pola dan karakteristik yang membedakan cyberbullying dari non-cyberbullying. Selama training, model akan mempelajari parameter weights melalui proses Backpropagation dan Gradient descent optimization berdasarkan data ini.
Sebagian kecil data, sekitar 10% atau sekitar 2.189 tweet, dialokasikan sebagai Data Validasi (Validation Data).

Data validasi memiliki peran krusial yang digunakan selama proses pelatihan untuk memantau performa model di setiap iterasi (epoch) secara real-time. Data ini tidak digunakan untuk updating model weights, melainkan untuk monitoring apakah model Overfitting atau Underfitting, serta membantu dalam penyesuaian Hyperparameter dan implementasi Early stopping mechanism untuk mencegah Overfitting. Validation performance menjadi Guidance untuk menentukan kapan training harus dihentikan dan konfigurasi mana yang optimal.
Sisa data, sekitar 10% atau sekitar 2.736 tweet, disimpan sebagai Data Uji (Test Data). Ini adalah aspek yang sangat penting: data uji sama sekali tidak pernah dilihat oleh model selama proses pelatihan dan validasi, sehingga merupakan Completely unseen data. Data uji hanya digunakan pada tahap evaluasi akhir setelah model selesai dilatih untuk mengukur kemampuan generalisasi model secara objektif dan unbiased. Test performance merupakan indikator final yang paling reliable tentang seberapa baik model akan perform pada data real-world yang baru.

E. Model Yang Diusulkan
Pelatihan model dalam penelitian ini dilakukan dengan melakukan Fine-tuning pada model IndoBERTweet, sebuah Pre-trained language model yang powerful. IndoBERTweet merupakan adaptasi dari arsitektur BERT (Bidirectional Encoder Representations from Transformers) yang telah dilatih secara khusus (Pre-trained) pada korpus data Twitter berbahasa Indonesia yang masif. Keunggulan IndoBERTweet adalah bahwa model ini telah memiliki pemahaman mendalam tentang konteks dan nuansa bahasa Indonesia informal, termasuk slang, abbreviations, dan Code-mixing yang prevalent di media sosial, sehingga penelitian ini tidak perlu melatih model dari nol (From scratch) yang akan memerlukan computational resources yang jauh lebih besar dan waktu yang lebih lama. Alur proses Fine-tuning ini diilustrasikan pada Gambar 2 yang menunjukkan arsitektur end-to-end dari input hingga output.

Gambar 2. IndoBERTweet Layer

Gambar 2. IndoBERTweet Layer

Proses dimulai ketika kalimat masukan (Input Sentence) dari Data Latih dimasukkan ke dalam alur proses. Kalimat tersebut terlebih dahulu dipecah menjadi serangkaian token melalui tahap Tokenization menggunakan tokenizer IndoBERTweet yang spesifik. Tokenizer ini menggunakan WordPiece tokenization yang mampu menangani out-of-vocabulary words dengan memecahnya menjadi subword units, sehingga sangat efektif untuk Handling informal language dan creative spellings yang umum di Twitter. Special tokens seperti [CLS] (untuk classification) ditambahkan di awal sequence dan [SEP] (separator) di akhir.
Token sequence selanjutnya diproses oleh arsitektur IndoBERTweet yang terdiri dari multiple layers. Di dalam arsitektur ini, token pertama-tama melewati lapisan Word embedding yang mengkonversi setiap token menjadi dense vector representation dalam high-dimensional space. Embedding layer ini tidak hanya mengandung Word embeddings, tetapi juga positional embeddings untuk menangkap informasi posisi token dalam sequence, dan segment embeddings untuk membedakan berbagai segments jika ada.

Setelah embedding, representasi ini kemudian diproses melalui multiple layers of Transformer Encoder yang merupakan core dari arsitektur BERT. Setiap Transformer layer terdiri dari multi-head self-Attention mechanism dan feed-forward neural networks. Self-Attention mechanism memungkinkan model untuk menangkap contextual relationships antar token dalam sequence secara bidirectional, sehingga setiap token representation dipengaruhi oleh semua token lainnya dalam context.

Ini sangat powerful untuk understanding nuances dan implicit meanings yang penting dalam cyberbullying detection.
Output dari final Transformer encoder layer kemudian diteruskan ke Dense Layer yang berfungsi sebagai Classification head.

Untuk task klasifikasi biner (cyberbullying vs non-cyberbullying),

classification layer menggunakan representasi dari [CLS] token yang telah mengagregasi informasi dari entire sequence, dan memetakannya ke dua output classes melalui fully connected layer dengan softmax activation.
Proses Fine-tuning sendiri terjadi dengan melatih kembali (updating) bobot (weights) pada beberapa lapisan terakhir dari IndoBERTweet, atau dalam beberapa kasus melakukan full Fine-tuning pada semua layers. Tujuannya adalah untuk meminimalkan error antara hasil prediksi dengan label sentimen yang sebenarnya (Ground truth). Loss function yang digunakan adalah Binary cross-entropy loss yang appropriate untuk binary classification task. Prosesnya dioptimalkan dengan mengatur Hyperparameter krusial seperti learning rate yang menentukan seberapa besar weight updates dilakukan, Batch size yang mempengaruhi gradient estimation stability, dan jumlah epoch yaitu berapa kali model melihat entire training dataset.
Selain arsitektur IndoBERTweet standalone, penelitian ini juga mengeksplorasi augmented architectures dengan menambahkan sequential layers setelah IndoBERTweet. Specifically, IndoBERTweet + BiLSTM dan IndoBERTweet + BiGRU architectures dikembangkan di mana output dari IndoBERTweet diteruskan ke Bidirectional LSTM atau Bidirectional GRU layers sebelum final classification layer. Sequential layers ini dapat menangkap additional temporal dependencies dan patterns yang mungkin bermanfaat untuk task ini.

F. Skenario Pengujian
Penelitian ini merancang serangkaian skenario pengujian yang comprehensive dan systematic untuk mengevaluasi berbagai aspek yang mempengaruhi performa model. Eksperimen dibagi menjadi dua tahap utama dengan tujuan yang berbeda.
Preliminary experiments dirancang untuk menentukan baseline performance dan memahami pengaruh dasar dari Hyperparameter utama. Tiga skenario diuji dengan fokus pada variasi learning rate dan Batch size, yang merupakan dua Hyperparameter paling influential dalam training deep neural networks.

Skenario pertama menggunakan konfigurasi dasar sebagai baseline: learning rate sebesar 2e-5 dengan Batch size 16. Pemilihan learning rate 2e-5 didasarkan pada recommendation dari literatur BERT Fine-tuning yang menunjukkan bahwa learning rate pada range 2e-5 hingga 5e-5 typically works well untuk most NLP tasks. Batch size 16 dipilih sebagai starting point yang balance antara gradient estimation quality dan memory efficiency.
Skenario kedua dirancang untuk menguji pengaruh learning rate yang sedikit lebih besar dengan menaikkannya menjadi 3e-5, sementara Batch size tetap dipertahankan pada 16.

Tujuannya adalah untuk mengobservasi apakah learning rate yang lebih tinggi dapat mempercepat Convergence atau justru menyebabkan instability dalam training process.
Skenario ketiga bertujuan untuk menguji pengaruh Batch size yang lebih besar dengan menaikkannya menjadi 32, sementara learning rate dikembalikan ke nilai baseline 2e-5. Batch size yang lebih besar dapat memberikan gradient estimates yang lebih stable dan accurate, namun memerlukan memory yang lebih besar dan mungkin mempengaruhi generalization.

Semua skenario dalam preliminary experiments dijalankan dengan jumlah epoch yang sama, yaitu 50 epoch, menggunakan set Data Latih dan Data Validasi yang identik untuk memastikan comparability. Hasil dari setiap skenario kemudian dibandingkan menggunakan validation metrics untuk menentukan konfigurasi dasar yang paling promising untuk investigasi lebih lanjut.
Berdasarkan insights dari preliminary experiments, tahap kedua melakukan investigasi yang lebih comprehensive dan systematic terhadap berbagai faktor yang mempengaruhi performa.

Optimizer and Learning rate scheduler Experiments: Eksperimen ini menguji pengaruh pemilihan Optimizer dan Learning rate scheduler terhadap Convergence dan final performance. Tiga algoritma optimasi diuji: Adam sebagai baseline Optimizer yang paling commonly used, AdamW (Adam with decoupled Weight decay) yang implements weight decay separately untuk regularization yang lebih efektif, dan SGD (Stochastic Gradient descent) sebagai representasi dari simpler non-adaptive Optimizer.

Setiap Optimizer dikombinasikan dengan dua tipe Learning rate scheduler: Linear scheduler yang menurunkan learning rate secara linear dari initial value ke 0 sepanjang training, dan Cosine Annealing scheduler yang menggunakan cosine function untuk smooth decay dengan kemampuan warm restarts. Kombinasi ini menghasilkan enam configurations yang berbeda untuk comprehensive comparison.
Architectural Comparison Experiments: Eksperimen ini membandingkan performance dari berbagai arsitektur model untuk mengidentifikasi optimal architecture.
Tujuh arsitektur diuji: (1) IndoBERTweet + BiGRU - combining Pre-trained model dengan Bidirectional GRU layers,
(2) IndoBERTweet + BiLSTM - combining dengan Bidirectional LSTM yang memiliki memory capacity lebih kompleks, (3) IndoBERTweet standalone - using Pre-trained model langsung tanpa additional layers, (4) CNN-LSTM hybrid - combining Convolutional Neural Networks untuk local feature extraction dengan LSTM untuk sequential modeling tanpa Pre-trained model, (5) BiGRU standalone - standalone sequential model tanpa Pre-training, (6) BiLSTM standalone - standalone sequential model untuk comparison, dan (7) SVM dengan TF-IDF - traditional Machine learning baseline untuk demonstrating value of Deep learning approaches.

Model-model dengan IndoBERTweet components dilatih menggunakan optimal configuration yang telah diidentifikasi dari Optimizer-scheduler experiments untuk ensuring fair comparison pada optimal conditions. Baseline models tanpa Pre-training menggunakan standard configurations dengan Adam Optimizer dan Linear scheduler. Semua models dievaluasi menggunakan identical test dataset untuk objective comparison.

G. Evaluasi Model
Model yang telah dilatih dievaluasi performanya secara menyeluruh dan Rigorous. Evaluasi ini dilakukan dengan menguji model menggunakan Data Uji yang merupakan Completely unseen data yang belum pernah dilihat oleh model sebelumnya, untuk mengukur kemampuan generalisasi model secara objektif dan unbiased. Kinerja model dianalisis secara kualitatif dan kuantitatif menggunakan berbagai metrik standar untuk classification tasks.
Analisis kuantitatif dilakukan dengan membuat Confusion Matrix yang memberikan gambaran rinci dan comprehensive mengenai distribusi hasil prediksi. Confusion Matrix menampilkan empat categories: True Positives (TP) yaitu cyberbullying yang correctly predicted sebagai cyberbullying, True Negatives (TN) yaitu non-cyberbullying yang correctly predicted sebagai non-cyberbullying, False Positives (FP) yaitu non-cyberbullying yang incorrectly predicted sebagai cyberbullying, dan False Negatives (FN) yaitu cyberbullying yang incorrectly predicted sebagai non-cyberbullying atau missed oleh model. Dari Confusion Matrix ini, dihitung metrik-metrik evaluasi standar untuk tugas klasifikasi yang memberikan different perspectives tentang model performance:

Accuracy (tingkat ketepatan keseluruhan) Rumus pada (1) menghitung rasio prediksi benar (positif dan negatif) terhadap keseluruhan data.

(1)

Precision (ketepatan prediksi positif) Rumus (2) menghitung seberapa akurat model saat memprediksi kelas positif.

(2)

Recall (kemampuan model untuk menemukan kembali semua data) Rumus ini menghitung kemampuan model menemukan kembali data positif yang sebenarnya. Ditunjukkan pada (3).

(3)

F1-Score (rata-rata harmonik dari Precision dan Recall) Rumus pada (4) menghitung rata-rata harmonik antara Precision dan Recall, yang memberikan keseimbangan antara keduanya.

(4)

III. Hasil dan Pembahasan

Karakteristik Data dan Pengaturan Eksperimen

Dataset yang digunakan dalam penelitian ini terdiri dari 13.677 tweet berbahasa Indonesia dengan karakteristik Code-mixed Indonesia-Inggris yang dikumpulkan dari platform media sosial X. Dataset dilabeli secara manual ke dalam dua kelas:

cyberbullying (kelas 1) dan non-cyberbullying (kelas 0),

dengan distribusi yang relatif seimbang untuk meminimalkan bias klasifikasi. Karakteristik unik dari dataset ini adalah prevalensi tinggi penggunaan bahasa gaul, singkatan, emoticon, dan Code-mixing yang mencerminkan pola komunikasi autentik pengguna media sosial Indonesia, terutama kalangan muda. Dataset dibagi menggunakan stratified random sampling dengan proporsi

80% untuk data latih, 10% untuk data validasi, dan 10% untuk data

uji, memastikan distribusi kelas yang konsisten di setiap subset.

Eksperimen dilakukan menggunakan framework PyTorch dengan Pre-trained model IndoBERTweet yang di-Fine-tune untuk tugas klasifikasi biner. Konfigurasi hardware mencakup GPU NVIDIA untuk mempercepat proses training, dengan batch processing dan mixed Precision training untuk efisiensi komputasi. Semua model dilatih selama 50 epoch dengan Early stopping mechanism berdasarkan performa validation loss untuk mencegah Overfitting.

Hyperparameter utama yang dieksplorasi meliputi learning rate (2e-5, 3e-5), Batch size (16, 32), tipe Optimizer (Adam, AdamW, SGD), dan Learning rate scheduler (Linear,

Cosine Annealing). Dropout rate ditetapkan pada 0.5 untuk semua eksperimen sebagai regularisasi standar.

Eksperimen Awal Sebagai Dasar

Serangkaian eksperimen awal dilakukan untuk menentukan konfigurasi baseline dan memahami perilaku dasar model IndoBERTweet pada dataset Code-mixed Indonesia-Inggris. Tiga skenario diuji dengan variasi learning rate dan Batch size, sebagaimana ditunjukkan pada Tabel 3.

Skenario pertama menggunakan learning rate 2e-5 dengan Batch size 16 menghasilkan akurasi 90.86% dan F1-Score 90.86%, menunjukkan performa baseline yang solid. Visualisasi training validation loss pada Gambar 3 memperlihatkan konvergensi yang stabil dengan training loss yang menurun tajam pada epoch-epoch awal dan mencapai nilai mendekati nol setelah epoch ke-10, sementara validation loss menunjukkan fluktuasi di kisaran 0.65-0.90 setelah epoch ke-20, mengindikasikan terjadinya Overfitting. Confusion Matrix pada Gambar 3 menunjukkan bahwa model berhasil mengklasifikasikan 1239 sampel kelas 0 dan 1246 sampel kelas 1 dengan benar, sementara terjadi 111 False Positives dan 140 False Negatives, mengindikasikan bahwa model memiliki kecenderungan sedikit lebih konservatif dalam memprediksi kelas negatif (0).

Gambar 3. Training, Validation Loss dan Confusion Matrix Skenario Pertama

Gambar 3. Training, Validation Loss dan Confusion Matrix Skenario Pertama

Gambar 4. Training, Validation Loss dan Confusion Matrix Skenario Kedua

Gambar 4. Training, Validation Loss dan Confusion Matrix Skenario Kedua

Skenario kedua dengan learning rate yang lebih tinggi (3e-5) menghasilkan penurunan performa menjadi 90.83% akurasi, sebagaimana ditunjukkan pada Gambar 4. Grafik validation loss menunjukkan instabilitas yang lebih besar dengan fluktuasi yang lebih tinggi di kisaran 0.70-0.95 dan puncak anomali mencapai hampir 0.95 pada epoch ke-19, mengkonfirmasi bahwa

learning rate yang terlalu tinggi menyebabkan overshooting dalam optimization landscape. Confusion Matrix pada Gambar 4 menampilkan 1246 True Positives untuk kelas 0 dan 1248 True Positives untuk kelas 1, dengan 104 False Positives dan 138 False Negatives, mengindikasikan bahwa learning rate yang tinggi menyebabkan model menghasilkan distribusi error yang sedikit berbeda namun dengan performa keseluruhan yang hampir serupa dengan skenario pertama.
Skenario ketiga dengan Batch size yang lebih besar (32) mencapai akurasi tertinggi 91.15% dan F1-Score 91.16%, dengan stabilitas konvergensi yang lebih baik seperti terlihat pada Gambar 5. Training loss menunjukkan pola penurunan yang smooth dan validation loss relatif lebih stabil di kisaran 0.70-0.90 dengan fluktuasi yang lebih terkontrol dibandingkan dua skenario sebelumnya. Confusion Matrix pada Gambar 5 menunjukkan peningkatan performa dengan 1214 True Positives untuk kelas 0 dan 1272 True Positives untuk kelas 1, dengan 136 False Positives dan 114 False Negatives, mengindikasikan bahwa Batch size yang lebih besar menghasilkan model yang lebih baik dalam mengidentifikasi kelas positif (1) dengan mengurangi False Negatives secara signifikan.

Gambar 5. Training, Validation Loss dan Confusion Matrix Skenario Ketiga

Gambar 5. Training, Validation Loss dan Confusion Matrix Skenario Ketiga

Tabel 3. Hasil Performa Dan Evaluasi Ekperimen Awal
Accuracy F1 Score
Skenario Pertama 90.86% 90.86%
Skenario Kedua 90.83% 90.83%
Skenario Ketiga 91.15% 91.16%

Tabel 3. Hasil Performa Dan Evaluasi Ekperimen Awal
Accuracy F1 Score
Skenario Pertama 90.86% 90.86%
Skenario Kedua 90.83% 90.83%
Skenario Ketiga 91.15% 91.16%

Analisis Optimizer dan Penjadwal Laju Pembelajaran
Investigasi sistematis terhadap pengaruh Optimizer dan Learning rate scheduler dilakukan untuk mengidentifikasi konfigurasi optimal dalam training IndoBERTweet untuk deteksi cyberbullying. Tiga algoritma optimasi diuji: Adam sebagai baseline Optimizer yang paling umum dalam NLP tasks,

AdamW dengan weight decay regularization yang terpisah,

dan SGD sebagai representasi dari non-adaptive Optimizer. Setiap Optimizer dikombinasikan dengan dua Learning rate scheduler: Linear scheduler yang menurunkan learning rate secara linear, dan Cosine Annealing scheduler yang menggunakan fungsi cosinus untuk penurunan learning rate yang lebih smooth dengan kemampuan untuk melakukan warm restart.
Hasil eksperimen Optimizer-scheduler ditunjukkan pada Tabel 4, mengungkapkan temuan yang signifikan. Kombinasi AdamW dengan Cosine Annealing scheduler dan Adam dengan Cosine Annealing scheduler keduanya mencapai performa tertinggi dengan akurasi 91.19%, Precision 91.5%, dan Recall 91%.

Hasil identikini menunjukkan bahwa untuk tugas deteksi cyberbullying pada dataset ini, Learning rate scheduler memiliki pengaruh yang lebih dominan dibandingkan perbedaan antara Adam dan AdamW. Cosine Annealing scheduler menghasilkan konvergensi yang lebih stabil dan mampu menemukan local minima yang lebih baik dibandingkan Linear scheduler, kemungkinan karena penurunan learning rate yang gradual memungkinkan model untuk melakukan Fine-tuning yang lebih halus pada tahap akhir training.
Kombinasi Adam dengan Linear scheduler dan AdamW dengan Linear scheduler menunjukkan performa yang sedikit lebih rendah dengan akurasi 90.86%. Meskipun perbedaan 0.33% terlihat modest, perbedaan ini konsisten dan menunjukkan bahwa Cosine Annealing scheduler memberikan keuntungan yang stabil dalam optimization process. Sebaliknya, kombinasi SGD dengan kedua scheduler (Cosine Annealing: 71.97%, Linear: 71.93%) menghasilkan performa yang jauh lebih rendah, dengan gap sebesar 19.22% dibandingkan adaptive Optimizers.

Hasil ini mengkonfirmasi bahwa adaptive learning rate per parameter yang ditawarkan oleh Adam dan AdamW sangat krusial untuk training architecture transformer yang kompleks seperti IndoBERTweet, di mana berbagai layer memiliki karakteristik gradient yang berbeda-beda.
Temuan ini konsisten dengan penelitian terdahulu pada transformer-based models yang menunjukkan superioritas adaptive Optimizers untuk Pre-trained language models. Analisis lebih lanjut mengungkapkan bahwa SGD gagal karena ketidakmampuannya untuk menyesuaikan learning rate secara individual untuk setiap parameter, yang sangat penting dalam Fine-tuning model Pre-trained di mana berbagai layer memerlukan update rate yang berbeda. Berdasarkan hasil ini, konfigurasi AdamW dengan Cosine Annealing scheduler dipilih untuk semua eksperimen arsitektur selanjutnya, dengan AdamW diprioritaskan karena implementasi weight decay yang terpisah yang lebih efektif dalam regularisasi.

Tabel 4. Hasil Performa Dan Evaluasi Analisis Optimizer
Optimazer Scheduler Accuracy Precision Recall
Adam Linear 90.

86% 91% 91%
Cosine 91.19% 91.5% 91%
AdamW Linear 90.86% 91% 91%

Cosine 91.19% 91.5% 91%
SGD Linear 71.93% 72% 72%
Cosine 71.

97% 72% 72%

Tabel 4. Hasil Performa Dan Evaluasi Analisis Optimizer
Optimazer Scheduler Accuracy Precision Recall
Adam Linear 90.

86% 91% 91%
Cosine 91.19% 91.5% 91%
AdamW Linear 90.86% 91% 91%
Cosine 91.19% 91.5% 91%
SGD Linear 71.93% 72% 72%
Cosine 71.

97% 72% 72%

Perbandingan Arsitektur dan Analisis Komponen
Untuk mengevaluasi kontribusi berbagai komponen arsitektur terhadap performa deteksi cyberbullying, penelitian ini melakukan perbandingan sistematis terhadap tujuh arsitektur yang berbeda, mencakup variasi IndoBERTweet dengan sequential layers,

baseline Deep learning models, dan traditional Machine learning approach.

Semua model berbasis IndoBERTweet dilatih menggunakan konfigurasi optimal yang telah diidentifikasi (AdamW Optimizer dengan Cosine Annealing scheduler), sementara baseline models menggunakan konfigurasi standar untuk memberikan perbandingan yang fair.
Hasil perbandingan arsitektur ditunjukkan pada Tabel 5. Model IndoBERTweet + BiLSTM mencapai performa terbaik dengan akurasi 91.48%, Precision 91.48%, Recall 91.48%, dan F1-Score 91.48%.

Superioritas arsitektur ini menunjukkan bahwa kombinasi Pre-trained language model dengan bidirectional sequential modeling layer mampu mengintegrasikan pemahaman konteks semantik yang kaya dengan kemampuan menangkap dependensi temporal dalam teks. BiLSTM layer memberikan kontribusi dalam modeling long-range dependencies yang penting untuk memahami nuansa cyberbullying yang sering tersebar di sepanjang tweet, seperti sarcasm yang memerlukan pemahaman konteks kalimat secara keseluruhan.

Model IndoBERTweet + BiGRU menunjukkan performa yang sangat kompetitif dengan akurasi 91.19%, Precision 91.22%, dan Recall 91.21%, hanya 0.29% lebih rendah dalam akurasi dibandingkan IndoBERTweet + BiLSTM. Perbedaan yang minimal ini menunjukkan bahwa BiGRU dapat menjadi alternatif yang efisien dengan keuntungan computational efficiency yang signifikan, mengingat BiGRU memiliki fewer parameters dan faster training time dibandingkan BiLSTM.

Untuk aplikasi real-time atau deployment dengan resource constraints, IndoBERTweet + BiGRU merupakan pilihan yang pragmatis tanpa mengorbankan performa secara signifikan. Secara menarik, model IndoBERTweet standalone mencapai akurasi 91.19% dengan Precision dan Recall 91%, menunjukkan performa yang sangat baik meskipun tanpa sequential layers tambahan. Hasil ini mengindikasikan bahwa Pre-trained model saja sudah memiliki kapasitas yang sangat baik dalam menangkap karakteristik cyberbullying melalui contextual embeddings yang kaya.

Fine-tuning IndoBE

RTweet saja sudah sufficient untuk banyak kasus,

dan penambahan sequential layers memberikan marginal improvement yang mungkin tidak selalu Justified mengingat tambahan kompleksitas dan computational cost. Namun, performa sedikit lebih tinggi dari IndoBERTweet + BiLSTM menunjukkan bahwa untuk achieving state-of-the-art results, sequential layers masih memberikan Value-add yang penting.
Model CNN-LSTM hybrid mencapai akurasi 88.

49% dengan Precision 89% dan Recall 88%,

menunjukkan gap 3% dibandingkan IndoBERTweet + BiLSTM. Meskipun CNN-LSTM mampu mengekstrak local features melalui convolutional layers dan menangkap sequential patterns melalui LSTM, arsitektur ini kurang optimal dibandingkan transformer-based approaches karena keterbatasan dalam menangkap long-range dependencies dan kurangnya Pre-trained knowledge tentang bahasa Indonesia informal. Hasil ini mengkonfirmasi bahwa Pre-trained contextual embeddings dari IndoBERTweet memberikan keuntungan yang substansial dibandingkan random initialized embeddings yang dilatih From scratch.
BiLSTM standalone mencapai akurasi 88.78% dengan Precision dan Recall 88%, sementara BiGRU standalone menghasilkan akurasi 87.39% dengan Precision dan Recall 87%, keduanya secara signifikan lebih rendah dibandingkan dengan versi yang diaugmentasi dengan IndoBERTweet. Perbedaan performa antara BiLSTM standalone (88.78%) dan IndoBERTweet + BiLSTM (91.48%) sebesar 2.7% secara jelas mendemonstrasikan kontribusi krusial dari Pre-trained language model. Lebih mencolok lagi, gap antara BiGRU standalone (87.39%) dan IndoBERTweet + BiGRU (91.19%) mencapai 3.8%, menekankan bahwa pemahaman konteks bahasa yang mendalam dari Pre-training pada data Twitter berbahasa Indonesia sangat penting untuk menangani kompleksitas linguistik dari Code-mixed text dan subtle nuances dari cyberbullying yang tidak explicit.

Perbandingan dengan Model Dasar Machine learning Tradisional
Untuk mendemonstrasikan keunggulan pendekatan Deep learning dan Pre-trained language models, penelitian ini mengimplementasikan Support Vector Machine (SVM) dengan ekstraksi fitur TF-IDF sebagai baseline Machine learning tradisional. SVM dipilih karena memiliki rekam jejak yang baik dalam tugas klasifikasi teks serta umum digunakan sebagai representasi pendekatan berbasis feature engineering pada penelitian deteksi perundungan di media sosial [27].

Model SVM dilatih menggunakan linear kernel dengan Hyperparameter C=1.0, yang sesuai untuk karakteristik fitur TF-IDF yang bersifat berdimensi tinggi dan sparse.
TF-IDF features diekstrak menggunakan vocabulary size 5000 dengan n-gram range (1,
2) untuk menangkap unigram dan bigram patterns,
minimum document frequency 2 untuk noise reduction, dan maximum document frequency 0.
95 untuk mengeliminasi terms yang terlalu umum. Sublinear TF scaling diterapkan untuk mengurangi efek dari term frequency yang sangat tinggi, dan parameter class_weight='balanced' digunakan untuk menangani ketidakseimbangan kelas dalam dataset.
Hasil evaluasi menunjukkan bahwa SVM + TF-IDF mencapai akurasi 78.18%, Precision 82.50%, Recall 78.42%, dan F1-Score 77.53%. Performa ini secara signifikan lebih rendah dibandingkan semua Deep learning models yang diuji, dengan gap terbesar mencapai 13.3% dibandingkan model terbaik (IndoBERTweet + BiLSTM: 91.48%). Perbedaan performa yang substansial ini mengungkapkan keterbatasan fundamental dari frequency-based features dalam menangkap kompleksitas semantik dan pragmatik dari cyberbullying dalam Code-mixed text. TF-IDF hanya mengandalkan statistical co-occurrence patterns dan tidak mampu memahami contextual meanings, figurative language, sarcasm, atau implicit threats yang merupakan karakteristik common dari cyberbullying di media sosial.
Menariknya, meskipun SVM menunjukkan Precision yang relatif tinggi (82.50%), Recall-nya jauh lebih rendah (78.42%), mengindikasikan bahwa model ini cenderung konservatif dalam memprediksi cyberbullying dan sering menghasilkan False Negatives.

Hal ini berbahaya dalam konteks aplikasi real-world karena banyak kasus cyberbullying yang tidak terdeteksi. Sebaliknya, model IndoBERTweet + BiLSTM menunjukkan balance yang baik antara Precision dan Recall (keduanya 91.48%), mengindikasikan kemampuan yang lebih Robust dalam mengidentifikasi cyberbullying tanpa menghasilkan terlalu banyak false alarms. Gap antara Precision dan Recall pada SVM (4.08%) menunjukkan trade-off yang kurang optimal, di mana model terlalu hati-hati dalam mengklasifikasikan teks sebagai cyberbullying, kemungkinan karena ketergantungan pada explicit keywords yang terbatas.

Analisis lebih mendalam mengungkapkan bahwa SVM mengalami kesulitan particular dalam mengidentifikasi cyberbullying yang subtle atau implicit, di mana harmful intent tidak diekspresikan melalui explicit hate words tetapi melalui konteks, tone, atau sarcasm. Misalnya, kalimat seperti
"wah pinter banget ya sampe segitunya"
dapat bermakna pujian atau ejekan tergantung konteks, yang tidak dapat ditangkap oleh bag-of-words representations bahkan dengan bigram features. Linear kernel yang digunakan hanya mampu melakukan linear separation pada feature space, yang tidak cukup untuk menangkap kompleksitas interaksi antar-kata yang non-linear dalam menentukan cyberbullying intent. Sebaliknya, IndoBERTweet dengan contextual embeddings-nya mampu membedakan nuansa ini berdasarkan pemahaman yang lebih holistik terhadap kalimat dan konteks komunikatif melalui Attention mechanisms yang memodelkan dependensi antar-kata secara dinamis.
Selain itu, model SVM menunjukkan limitasi dalam Handling Code-mixed text karena feature engineering tradisional tidak dirancang untuk menangani linguistic code-switching. Meskipun bigram features dapat menangkap beberapa kombinasi kata lintas-bahasa, representasi ini tetap terlalu shallow untuk memahami semantik dari Code-mixing yang natural terjadi dalam komunikasi Indonesia-Inggris.
Pre-trained models seperti IndoBERTweet,
yang dilatih pada data Twitter yang naturally mengandung Code-mixing, memiliki inherent capability untuk memproses dan memahami mixed-language expressions tanpa memerlukan explicit Handling mechanisms.

Gap performa sebesar 13.3% bukan hanya menunjukkan keunggulan quantitative dari Deep learning approaches, tetapi juga mengindikasikan perbedaan qualitative dalam bagaimana kedua paradigma ini merepresentasikan dan memahami language. Linear kernel yang efisien untuk high-dimensional sparse data tetap tidak mampu mengatasi keterbatasan fundamental dari TF-IDF features yang tidak memiliki semantic understanding.

 Tabel 5. Hasil Performa Dan Evaluasi Perbandingan Model
Model Accuracy Precision Recall F1-Score
IndoBERTweet + BiLSTM 91.

48% 91.48% 91.48% 91.48%
IndoBERTweet + BiGRU 91.19% 91.22% 91.21% 91.19%
IndoBERTweet 91.19% 91% 91% 91.19%
CNN-LSTM Hybrid 88.49% 89% 88% 88%
BiLSTM 88.78% 88% 88% 88.78%
BiGRU 87.39% 87% 87% 87.39%
SVM + TF-IDF 78.18% 82.50% 78.42% 77.53%

Tabel 5.

Hasil Performa Dan Evaluasi Perbandingan Model
Model Accuracy Precision Recall F1-Score
IndoBERTweet + BiLSTM 91.

48% 91.48% 91.48% 91.48%
IndoBERTweet + BiGRU 91.19% 91.22% 91.21% 91.19%
IndoBERTweet 91.19% 91% 91% 91.19%
CNN-LSTM Hybrid 88.49% 89% 88% 88%

BiLSTM 88.78% 88% 88% 88.78%
BiGRU 87.39% 87% 87% 87.39%
SVM + TF-IDF 78.

18% 82.50% 78.42% 77.53%

Analisis Kesalahan dan Interpretasi Model
Untuk memahami lebih mendalam kekuatan dan keterbatasan model terbaik (IndoBERTweet + BiLSTM), analisis error dilakukan menggunakan Confusion Matrix pada test dataset Gambar 6.

Confusion Matrix pada Gambar 6 menunjukkan bahwa dari 1.350 sampel kelas non-cyberbullying (kelas 0), model berhasil mengklasifikasikan 1.234 sampel dengan benar (True Negatives), namun terdapat 116 sampel yang salah diprediksi sebagai cyberbullying (False Positives). Sedangkan dari 1.386 sampel kelas cyberbullying (kelas 1), model berhasil mengklasifikasikan 1.269 sampel dengan benar (True Positives), dengan 117 sampel yang salah diprediksi sebagai non-cyberbullying (False Negatives).

Hasil ini menunjukkan bahwa model memiliki performa yang seimbang dengan tingkat kesalahan yang relatif rendah pada kedua kelas (8.6% untuk False Positives dan 8.4% untuk False Negatives).

Gambar.

6. Confusion Matrix IndoBERTweet + BiLSTM

Gambar.

6. Confusion Matrix IndoBERTweet + BiLSTM

Analisis terhadap 116 False Positives (non-cyberbullying yang diprediksi sebagai cyberbullying) mengungkapkan bahwa sebagian besar kesalahan terjadi pada tweet yang mengandung bahasa yang keras atau emotionally charged namun tidak dalam konteks bullying, seperti kritik terhadap public figures atau ekspresi frustrasi personal.

Model cenderung menginterpretasikan aggressive language atau negative sentiment sebagai cyberbullying meskipun tidak ditargetkan kepada individu tertentu untuk tujuan harassment.
Contohnya, tweet seperti
"pemerintah ini beneran ngeselin banget sih,
bikin kesel aja!"
mengandung kata-kata negatif yang kuat namun merupakan kritik terhadap institusi, bukan personal attack. Hal ini menunjukkan bahwa meskipun model mampu mendeteksi toxic language dengan baik, pemahaman tentang intentionality dan target specificity masih menjadi challenge.

117 False Negatives (cyberbullying yang diprediksi sebagai non-cyberbullying) predominantly terjadi pada kasus-kasus yang melibatkan sarcasm yang sangat subtle, indirect bullying yang menggunakan euphemism, atau cyberbullying yang embedded dalam konteks percakapan yang lebih luas yang tidak tercakup dalam single tweet. Misalnya, tweet yang mengatakan "semoga cepet sadar ya" atau "wah pinter banget deh kamu" tanpa konteks additional mungkin terlihat innocuous atau bahkan seperti pujian, namun dalam konteks harassment campaign bisa merupakan subtle bullying dengan tone sarkastik.

Keterbatasan ini inherent pada tweet-level classification dan menunjukkan bahwa untuk truly comprehensive cyberbullying detection, conversation-level atau thread-level analysis mungkin diperlukan untuk menangkap konteks yang lebih luas dari interaksi sosial.
Menariknya, dari total 2.736 sampel test, model mencapai akurasi 91.48% dengan 2.503 prediksi benar dan hanya 233 kesalahan (116 FP + 117 FN). Distribusi error yang hampir seimbang antara False Positives dan False Negatives (rasio 116:117 atau hampir 1:1) menunjukkan bahwa model tidak memiliki bias yang signifikan terhadap salah satu kelas, yang merupakan karakteristik penting untuk aplikasi praktis. Model menunjukkan performa yang baik dalam menangani Code-mixing, dengan error rate yang tidak secara disproportionate tinggi pada fully Code-mixed instances dibandingkan dengan monolingual Indonesian tweets. Hal ini mengkonfirmasi efektivitas IndoBERTweet's Pre-training dalam menangani variasi linguistik yang umum di media sosial Indonesia. Namun, model masih mengalami kesulitan dengan highly informal language,
excessive abbreviations (seperti "gk", "yg", "bgt"), atau creative spelling variations yang sangat jauh dari standard forms (seperti "syantiiikk",

"gemes bgt sih lohh"
), menunjukkan bahwa continuous learning atau periodic retraining dengan data terbaru mungkin diperlukan untuk maintaining performance seiring evolusi bahasa di media sosial.

Ringkasan Kinerja Komprehensif dan Wawasan
Tabel 5 merangkum performa keseluruhan dari semua model yang dievaluasi dalam penelitian ini, memberikan pandangan holistik tentang landscape berbagai approaches untuk deteksi cyberbullying pada Code-mixed Indonesian-English text. Hasil menunjukkan hierarki yang jelas: model berbasis IndoBERTweet dengan sequential layers di tier tertinggi (91.19-91.48%), model berbasis IndoBERTweet standalone di tier kedua (91.19%), baseline Deep learning models di tier ketiga (87.39-88.78%), dan traditional Machine learning di tier terendah (78.18%).

Analisis komprehensif mengungkapkan beberapa insights kunci yang memiliki implikasi both theoretical dan practical. Pertama, kontribusi Pre-trained language model terbukti sangat substantial, dengan peningkatan akurasi berkisar 2.7-3.8% ketika IndoBERTweet ditambahkan ke sequential layers.
Kontribusi ini bukan hanya dalam bentuk quantitative improvement,
tetapi qualitative enhancement dalam model's ability to understand subtle linguistic phenomena.

Kedua, sequential layers (BiLSTM/BiGRU) memberikan marginal yet consistent improvement,

dengan peningkatan 0.29% dari IndoBERTweet standalone,

suggesting bahwa untuk resource-constrained scenarios, IndoBERTweet standalone sudah sufficient, namun untuk achieving maximum performance, sequential layers masih valuable.

Ketiga, pemilihan Optimizer memiliki impact yang drastically significant,

dengan gap 19.22% antara adaptive Optimizers (Adam/AdamW) dan non-adaptive Optimizer (SGD).

Temuan ini menekankan bahwa algorithmic choices dalam optimization process sangat critical untuk success of transformer-based models. Keempat, Learning rate scheduler memberikan consistent yet modest improvement (0.33%), suggesting bahwa while important, scheduler choice adalah secondary compared to Optimizer choice. Kelima, gap signifikan 13.3% antara best Deep learning model dan SVM baseline secara convincing demonstrates paradigm shift dari feature engineering ke representation learning dalam NLP.

Dari perspective praktis untuk deployment, hasil penelitian ini menunjukkan bahwa IndoBERTweet + BiLSTM dengan AdamW Optimizer dan Cosine Annealing scheduler merupakan optimal choice ketika maximum Accuracy adalah prioritas.

Namun, untuk scenarios di mana computational efficiency penting atau real-time processing diperlukan,

IndoBERTweet + BiGRU atau bahkan IndoBERTweet standalone dapat menjadi alternatif yang pragmatis dengan trade-off yang minimal dalam performa. Understanding trade-offs ini penting untuk practical implementation dalam production environments di mana factors seperti latency, throughput, dan resource consumption harus dipertimbangkan alongside Accuracy.

Perbandingan dengan Teknologi Terkini dan Karya Terkait

Perbandingan dengan penelitian terdahulu menunjukkan bahwa model IndoBERTweet + BiLSTM yang dikembangkan dalam penelitian ini (91.48% akurasi) mengungguli sebagian besar approaches existing untuk cyberbullying detection pada teks berbahasa Indonesia. Zakaria, Nurjannah, dan Nurrahmi [25] yang menggunakan IndoBERTweet untuk deteksi misogyny pada TikTok mencapai akurasi 76.89%, significantly lower dibandingkan hasil penelitian ini.

Kusuma, dan Chawonda [26] dengan IndoBERTweet + BiLSTM untuk hate speech detection mencapai akurasi tertinggi 93.

7%, namun pada dataset yang berbeda dan task yang slightly different (hate speech vs cyberbullying), making direct comparison challenging.
Penelitian-penelitian yang menggunakan traditional Machine learning approaches menunjukkan performa yang considerably lower. Abdullah [12] dengan SVM mencapai akurasi 99.75%, namun hasil ini likely overstated karena potential issues dengan data imbalance yang tidak diaddress, as noted in penelitian tersebut.

Rizki [13] dengan SVM mencapai 70% dan Hasan [14] dengan KNN mencapai 71.43%, both substantially lower dibandingkan pendekatan Deep learning. Among Deep learning approaches, Widiyantoro [20] dengan LSTM standalone mencapai Accuracy sekitar 59.7% (improved through tuning), dan Andika [21] dengan CNN-LSTM mencapai F1-Score 0.84, both lower dibandingkan hasil penelitian ini.
Yang particularly notable adalah perbandingan dengan Rosid [22] yang menangani Code-mixed Indonesian-English text untuk sarcasm detection dengan CNN + multihead attention + BiGRU, mencapai 94.60%.

Meskipun Accuracy mereka slightly higher, task mereka (sarcasm detection) arguably simpler dibandingkan cyberbullying detection yang encompasses broader range of behaviors. Moreover, mereka melaporkan significant performance drop ke 88.

02% pada dataset berbeda, highlighting generalization challenges.

Sebaliknya, model dalam penelitian ini menunjukkan consistent performance dan Robust generalization pada test data yang unseen.

Keunggulan komparatif penelitian ini terletak pada several factors: (1) systematic Hyperparameter optimization yang comprehensive,

(2) explicit focus pada Code-mixed text yang prevalent di media sosial Indonesia namun often neglected dalam penelitian existing, (3) Rigorous comparison across multiple architectures dan traditional baselines untuk demonstrating clear value proposition dari approach yang diusulkan, dan (4) practical applicability dengan consideration terhadap efficiency trade-offs. Hasil penelitian ini advances state-of-the-art dalam cyberbullying detection untuk Indonesian context dan provides validated approach yang dapat diadopsi untuk practical deployment.

Keterbatasan dan Arah Penelitian Masa Depan

Meskipun penelitian ini mencapai results yang promising, beberapa keterbatasan perlu diakui untuk contextualizing temuan dan guiding future research. Pertama, dataset yang digunakan terbatas pada 13.677 tweets, yang meskipun adequate untuk demonstrating Proof-of-concept dan achieving good performance, masih relatively modest dalam scale. Larger datasets dapat potentially improve model's ability untuk generalize across diverse types dan subtle variations dari cyberbullying.

Additionally, dataset dikumpulkan dari media sosial X dan mungkin tidak fully representative dari cyberbullying di platforms lain seperti Instagram, TikTok, atau Facebook yang memiliki different user demographics dan communication norms.
Kedua, penelitian ini menggunakan tweet-level classification,
yang inherently limited dalam capturing contextual dependencies yang extend beyond single tweet. Cyberbullying sering terjadi dalam context dari conversation threads atau sustained harassment campaigns, dan tweet-level analysis mungkin miss important contextual cues. Future work dapat explore conversation-level atau thread-level classification untuk capturing broader context. Ketiga, model saat ini treats cyberbullying sebagai binary classification problem, padahal cyberbullying encompasses diverse types (harassment, doxing, sexual harassment, hate speech, etc.) dengan different characteristics dan severities. Multi-class classification atau hierarchical taxonomy dapat provide more nuanced detection yang useful untuk targeted interventions.
Keempat, meskipun model mencapai good quantitative performance,
penelitian ini belum fully address explainability dan interpretability.
Untuk practical deployment, terutama dalam high-stakes applications seperti content moderation, understanding why model makes certain predictions adalah crucial untuk trust dan accountability. Future research dapat incorporate attention visualization, LIME, SHAP, atau explainability techniques lainnya untuk providing transparent insights into model's decision-making process.

Kelima, penelitian ini conducted dalam controlled experimental setting dan belum validated dalam real-world deployment conditions di mana factors seperti data drift, adversarial attacks, atau evolving language patterns dapat impact performance over time.
Several promising directions untuk future research emerge dari keterbatasan ini. Integration dengan multimodal information (images, videos, user profiles) dapat enhance detection Accuracy, particularly untuk cases di mana text alone Ambiguous.

Development dari real-time detection systems dengan efficient architectures untuk on-the-fly content moderation dapat enable proactive interventions. Exploration dari few-shot learning atau transfer learning approaches dapat facilitate adaptation ke new platforms atau languages dengan minimal labeled data. Implementation dari federated learning dapat enable model training across multiple platforms sambil preserving user privacy. Finally, interdisciplinary collaboration dengan psychologists, sociologists, dan policy experts dapat ensure bahwa technical solutions effectively address social problem dari cyberbullying dalam culturally appropriate dan ethically sound manner.

VII. Kesimpulan

Penelitian ini berhasil mengembangkan sistem deteksi cyberbullying pada media sosial X menggunakan model IndoBERTweet yang telah dioptimalkan untuk teks Code-mixed Indonesia-Inggris. Hasil eksperimen menunjukkan bahwa model IndoBERTweet + BiLSTM dengan Optimizer AdamW dan Learning rate scheduler Cosine Annealing memberikan kinerja terbaik, dengan akurasi dan F1-Score sebesar 91,48%. Temuan ini menunjukkan bahwa penggunaan adaptive Optimizer seperti AdamW menghasilkan peningkatan yang signifikan dibandingkan dengan SGD, dengan perbedaan akurasi mencapai 19,22%.

Selain itu, kontribusi Pre-trained IndoBERTweet dalam meningkatkan performa juga terlihat jelas, dengan peningkatan akurasi sebesar 2,7% untuk arsitektur BiLSTM dan 3,8% untuk arsitektur BiGRU dibandingkan dengan model sequential layer tanpa Pre-training. Penelitian ini juga menunjukkan bahwa model berbasis Deep learning dapat menangani kompleksitas linguistik yang ada pada teks Code-mixed, yang merupakan tantangan utama dalam analisis sentimen pada platform media sosial. Perbandingan dengan metode Machine learning tradisional seperti SVM yang menggunakan TF-IDF sebagai fitur menunjukkan bahwa model Deep learning memberikan hasil yang jauh lebih baik, dengan perbedaan akurasi mencapai 13,30%. Temuan ini memberikan kontribusi penting dalam pengembangan sistem deteksi cyberbullying otomatis yang lebih efektif dan efisien, yang dapat diterapkan dalam sistem moderasi konten untuk platform media sosial. Meskipun demikian, penelitian ini memiliki beberapa keterbatasan, seperti penggunaan dataset yang terbatas pada satu platform dan belum diterapkannya explainable AI untuk menjelaskan keputusan model.

Penelitian selanjutnya dapat memperluas dataset ke platform lain, mengembangkan sistem deteksi multi-kelas untuk mengidentifikasi berbagai jenis cyberbullying, serta mengintegrasikan sistem deteksi ini untuk bekerja dalam deteksi waktu nyata guna meningkatkan efektivitas moderasi konten di media sosial.

Ucapan Terima Kasih

Penelitian ini mendapatkan dukungan keuangan dari Kementerian Pendidikan Tinggi, Ilmu Pengetahuan, dan Teknologi melalui Skema Penelitian Dasar Rutin (nomor hibah 128/C3/DT.05.00/PL/2025).

Referensi

[1]A.

N. Alfifa, V. Fahira, R. Batubara, K. Sakinah, J. Manik, and R. A.

Yoes, “Media

Sosial dan Pembentukan Opini Publik (Analisis Studi Kasus Echo Chamber Pada Interaksi Komentar di Akun Instagram @Turnbackhoaxid Dalam Konteks

Post-Truth),” J. Penelit. Ilmu-Ilmu Sos., vol. 2, no. 6, pp. 162–169, 2025, [Online]. Available: https://ojs.daarulhuda.or.id/index.php/Socius/article/view/1130
[2]S. Widi, “Pengguna Media Sosial Di Indonesia Sebanyak 167 Juta Pada 2023,” dataindonesia.id. [Online]. Available: https://dataindonesia.id/internet/detail/pengguna-media-sosial-diindonesia-sebanyak-167-juta-pada-2023
[3]F. A. Imani, A. Kusmawati, and H. M. T.

Amin,

“Pencegahan Kasus Cyberbullying Bagi Remaja Pengguna Sosial Media,” Khidm. Sos. J. Soc. Work Soc. Serv., vol. 2, no. 1, pp. 74–83, 2021, [Online]. Available: https://jurnal.umj.ac.id/index.php/khidmatsosial/article/view/10433
[4]J. M. Beaton, W. J. Doherty, and L.

M. Wenger,

“Mothers and fathers coparenting together,” in The Routledge Handbook of Family Communication, London: Routledge, 2012, pp. 237–252. [Online]. Available: https://www.taylorfrancis.com/chapters/edit/10.4324/9780203848166-22/mothers-fathers-coparenting-together-john-beaton-william-doherty-lisa-wenger
[5]A. Sukmawati,

A. Puput, and B. Kumala,

“Dampak Cyberbullying Pada Remaja Di Media Sosial,”

Alauddin Sci. J. Nurs., vol. 2020, no. 1, pp. 55–65, 2020,

[Online]. Available: http://journal.uin-alauddin.ac.id/index.php/asjn/issue/view/1328
[6]M. S. Jinan, M. R. Handayani, M. A. Ulinuha, and K.

Umam,

“Muhammad Syifaaul Jinan 1) , Maya Rini Handayani 2) , Masy Ari Ulinuha 3) , Khothibul Umam* 4),”

vol. 10, no. 3, pp. 2666–2678, 2025.
[7]Al-Khowarizmi,

I. P. Sari, and H. Maulana, “Detecting Cyberbullying on Social Media Using Support Vector Machine: A Case Study on Twitter,” Int.

J. Saf. Secur. Eng., vol. 13, no. 4, pp. 709–714, 2023, doi: 10.18280/ijsse.130413.
[8]A. Palagati,

S. K. Balan, S. Arun Joe Babulo, L. Raja, K. K. Natarajan, and R. Kalimuthu, “Comparative Analysis of Machine Learning Algorithms and Datasets for Detecting Cyberbullying on Social Media Platforms,” Int. Conf. Comput. Intell. Real.

Technol. Proc. ICCIRT 2024, pp. 391–396, 2024, doi: 10.1109/ICCIRT59484.

2024.10922033.
[9]N. Novalita, A. Herdiani, I. Lukmana, and D. Puspandari, “Cyberbullying identification on twitter using random forest classifier,” J. Phys. Conf. Ser., vol. 1192,

no. 1, 2019, doi: 10.1088/1742-6596/1192/1/012029.
[10]A. U. Rehman, A. K. Malik,

B. Raza, and W. Ali, “A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis,” Multimed. Tools Appl., vol. 78, no. 18, pp. 26597–26613, Sep. 2019, doi: 10.1007/s11042-019-07788-7.
[11]Y. Kim, “Convolutional neural networks for sentence classification,” EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc.

Conf., pp. 1746–1751, 2014, doi: 10.3115/v1/d14-1181.
[12]N. Abdulloh and A. F.

Hidayatullah,

“Deteksi Cyberbullying pada Cuitan Media Sosial Twitter,” Automata, vol. Vol 1, no. 1, pp. 1–5, 2021.
[13]M.

3 **dx.doi.org** | VISUALISASI ANALISIS SENTIMEN SIBERBULLYING PADA POST INSTAGRAM MENGGUNAKAN ORANGE DATA MINING
http://dx.doi.org/10.59095/ijcsr.v2i1.15

F. Rizki,
K. Auliasari, and R. Primaswara Prasetya,
“Analisis Sentiment Cyberbullying Pada Sosial Media Twitter Menggunakan Metode Support Vector Machine,”
JATI (Jurnal Mhs. Tek. Inform., vol. 5, no. 2, pp. 548–556, 2021, doi: 10.36040/jati.v5i2.

3808.
[14]N. F.

Hasan,

“Deteksi Cyberbullying pada Facebook Menggunakan Algoritma K-Nearest Neighbor,”

J. Smart Syst., vol. 1, no. 1, pp. 35–44, 2021, doi: 10.36728/jss.v1i1.1605.
[15]R. Masbadi Hatullah Nurnaryo, M. Mulaab, I. Oktavia Suzanti, D. Abdul Fatah, A. D. Cahyani, and F.

Ayu Mufarroha, “Deteksi

4 **dx.doi.org** | Deteksi Komentar Cyberbullying Pada YouTube Dengan Metode Convolutional Neural Network – Long Short-Term Memory Network (CNN-LSTM)
http://dx.doi.org/10.34148/teknika.v12i3.677

Cyberbullying Pada Data Tweet Menggunakan Metode Random Forest Dan Seleksi Fitur Information Gain,”
J. Simantec, vol. 11, no. 1, pp. 33–40, 2022, doi: 10.21107/simantec.v11i1.17256.
[16]H. Santoso, R.

A. Putri, and S. Sahbandi, “Deteksi Komentar Cyberbullying pada Media Sosial Instagram Menggunakan Algoritma Random Forest,”

J. Manaj. Inform., vol. 13, no. 1, pp. 62–72, 2023, doi: 10.34010/jamika.v13i1.9303.
[17]Fauzan

5 **dx.doi.org** | Perbandingan analisis sentimen pada aplikasi SIREKAP dengan aplikasi SITUNG di media sosial X menggunakan algoritma Support Vector Machine
http://dx.doi.org/10.33022/ijcs.v13i4.4084

Baehaqi and N.
Cahyono,
“Analisis Sentimen Terhadap Cyberbullying Pada Komentar Di Instagram Menggunakan Algoritma Naïve Bayes,”
Indones. J. Comput. Sci., vol. 13, no. 1, pp. 1051–1063, 2024, doi: 10.33022/ijcs.v13i1.3301.
[18]A. Machmud, B. Wibisono, and N.

Suryani, “Analisis Sentimen Cyberbullying Pada Komentar X Menggunakan Metode Naïve Bayes,” vol. 5, no. 1, 2025.
[19]R.

6 **doi.org** | Implementasi Data Mining Untuk Klasifikasi Komentar Hate Speech Menggunakan Algoritma Support Vector Machine (SVM)
https://doi.org/10.30591/smartcomp.v14i3.8122

Triyana, O. Virgantara Putra, and F. R. Pradhana,
“Deteksi Cyberbullying Pada Tweet Berbahasa Inggris Dengan Metode Support Vector Machine,”

Semin. Nas. Has. Penelit. Pengabdi.

Masy. Bid. Ilmu Komput., pp. 98–103, 2022.
[20]P.

7 **sistemasi.ftik.unisi.ac.id** | Applying Artificial Intelligence to Analyze Emotions in Social Media Comments using Large Language Models
https://sistemasi.ftik.unisi.ac.id/index.php/stmsi/article/download/5187/981

Widiyantoro and Y.
D. Prasetyo,
“Deteksi Cyberbullying pada Pemain Sepak Bola di Platform Media Sosial ‘ X ’ Menggunakan Metode Long Short-Term Memory (LSTM),
”
2025.
[21]A. J. Andika, Y. Kristian, and E. I.

Setiawan, “Deteksi

8 **dx.doi.org** | Deteksi Komentar Cyberbullying Pada YouTube Dengan Metode Convolutional Neural Network – Long Short-Term Memory Network (CNN-LSTM)
http://dx.doi.org/10.34148/teknika.v12i3.677

Komentar Cyberbullying Pada YouTube Dengan Metode Convolutional Neural Network – Long Short-Term Memory Network

(CNN-LSTM),” Teknika, vol. 12, no. 3, pp.

183–188, 2023, doi: 10.34148/teknika.v12i3.677.
[22]M. A. Rosid, D.

Siahaan, and A. Saikhu, “Sarcasm

9 **scholar.google.com** | Mochamad Alfan Rosid - Google Scholar
https://scholar.google.com/citations?user=JAN9FloAAAAJ&hl=id

Detection in Indonesian-English Code-Mixed Text Using Multihead Attention-Based Convolutional and Bi-Directional

GRU,”

IEEE Access, no. July, pp. 137063–137079, 2024, doi: 10.1109/ACCESS.2024.3436107.
[23]J.

10 **dx.doi.org** | Analisis Sentimen Review Hotel Menggunakan Metode Deep Learning BERT
http://dx.doi.org/10.24002/jbi.v14i02.7244

Devlin, M.
W. Chang, K. Lee, and K. Toutanova,
“BERT: Pre-training of deep bidirectional transformers for language understanding,”
NAACL HLT 2019 - 2019

Conf.

11 **eksplora.stikom-bali.ac.id** | Part of Speech Tagging Pada Teks Bahasa Indonesia dengan BiLSTM + CNN + CRF dan ELMo
https://eksplora.stikom-bali.ac.id/index.php/eksplora/article/download/506/204

North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, pp.

4171–4186,

2019.
[24]M. Safitri et al.,

“DETEKSI CYBERBULLYING TWEET MENGGUNAKAN MACHINE,”

pp. 370–374, 2025.
[25]P. H. Zakaria, D. Nurjannah, and H.

Nurrahmi, “Misogyny Text Detection on Tiktok Social Media in Indonesian Using the Pre-trained Language Model IndoBERTweet,” J. Media Inform. Budidarma, vol. 7, no. 3, p. 1297, 2023,

doi: 10.30865/mib.v7i3.6438.
[26]J. Forry Kusuma and A. Chowanda,

“

12 **journal.trunojoyo.ac.id**
https://journal.trunojoyo.ac.id/edutic/article/download/28822/10616

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION journal homepage : www.joiv.org/index.php/joiv INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

Indonesian Hate Speech Detection Using IndoBERTweet and BiLSTM on Twitter,”

vol. 7, no. September, pp. 773–780,

2023, [Online]. Available: www.joiv.org/index.php/joiv
[27]S. Nauli, S. S. Berutu, H. Budiati, and F.

Maedjaja,

“Klasifikasi Kalimat Perundungan Pada Twitter Menggunakan Algoritma Support Vector Machine,”

JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform., vol. 10, no. 1, pp. 107–122, 2025, doi: 10.29100/jipi.v10i1.

5749.