Hybrid CNN-LSTM for Indonesian Cyberbullying Text Classification on Social Media X

Muhammad Hafizh Fattah1), Mochamad Alfan Rosid2) *, Sukma Aji3) & Suprianto4)

1,2,3,4) Program Studi Informatika, Fakultas Sains dan Teknologi, Universitas Muhammadiyah Sidoarjo, Indonesia
* Coresponding Email: alfanrosid@umsida.ac.id

Abstract. *Cyberbullying on social media platform X has become a critical digital threat and requires automatic detection mechanisms to mitigate psychological impacts on victims. This study proposes a hybrid deep learning architecture that combines Convolutional Neural Network (CNN) for local feature extraction and Long Short-Term Memory (LSTM) for sequential context understanding in classifying Indonesian language cyberbullying comments. This study evaluates model performance using a dataset of 13,677 comments from social media X through a series of systematic testing scenarios, including the impact of regularization, utilization of FastText embeddings, and comparative studies against state-of-the-art models. Experimental results demonstrate that the Early Stopping mechanism is a critical factor in this architecture, where without this mechanism the model experiences accuracy degradation of up to 32%. The proposed CNN-LSTM model achieves 88.38% accuracy and 88.00% F1-Score, improving to 0.9559 AUC with FastText integration. This model achieves over 97% of IndoBERTweet's performance with 22 times lower computational complexity (4.97 million versus 110.88 million parameters) and outperforms machine learning methods such as SVM with an accuracy margin of more than 10 percentage points. This study concludes that the CNN-LSTM architecture offers a robust and efficient solution for cyberbullying detection, particularly for resource-constrained environments.*

Keywords - *Cyberbullying, CNN-LSTM, Deep Learning, Social Media X, Indonesian Language.;*

Abstrak. *Perundungan siber di platform media sosial X telah menjadi ancaman digital yang kritis dan memerlukan mekanisme deteksi otomatis untuk memitigasi dampak psikologis pada korban. Penelitian ini mengusulkan arsitektur deep learning hibrida yang menggabungkan Convolutional Neural Network (CNN) untuk ekstraksi fitur lokal dan Long Short-Term Memory (LSTM) untuk pemahaman konteks sekuensial dalam mengklasifikasikan komentar perundungan siber berbahasa Indonesia. Penelitian ini mengevaluasi kinerja model menggunakan dataset sebanyak 13.677 komentar dari media sosial X melalui serangkaian skenario pengujian sistematis, termasuk dampak regularisasi, pemanfaatan embedding FastText, dan studi komparatif terhadap model mutakhir. Hasil eksperimen menunjukkan bahwa mekanisme Early Stopping merupakan faktor kritis dalam arsitektur ini, di mana tanpa mekanisme tersebut model mengalami degradasi akurasi hingga 32%. Model CNN-LSTM yang diusulkan mencapai akurasi 88,38% dan F1-Score 88%, meningkat menjadi AUC 0,9559 dengan integrasi FastText. Model ini mencapai lebih dari 97% kinerja IndoBERTweet dengan kompleksitas komputasi 22 kali lipat lebih rendah (4,97 juta berbanding 110,88 juta parameter) dan mengungguli metode machine learning seperti SVM dengan margin akurasi lebih dari 10 poin persentase. Penelitian ini menyimpulkan bahwa arsitektur CNN-LSTM menawarkan solusi robust dan efisien untuk deteksi perundungan siber, khususnya untuk lingkungan dengan sumber daya terbatas.*

Kata Kunci - *Perundungan Siber, CNN-LSTM, Deep Learning, Media Sosial X, Bahasa Indonesia.*

I. PENDAHULUAN

Era digital kontemporer ditandai dengan penetrasi media sosial yang masif ke dalam berbagai aspek kehidupan masyarakat. Platform seperti X (sebelumnya dikenal sebagai Twitter) telah bertransformasi dari sekadar jejaring sosial menjadi arena utama untuk wacana publik, diseminasi informasi, dan interaksi sosial berskala global. Statistik terbaru dari tahun 2024 (Shubham Singh, 2025) menunjukkan bahwa X memiliki sekitar 611 juta pengguna aktif bulanan di seluruh dunia dan menempati peringkat di antara platform media sosial teratas secara global. Skala besar ini menjadikan X sebagai salah satu ekosistem komunikasi paling berpengaruh. Platform ini memainkan peran penting sebagai ruang untuk ekspresi pendapat yang cepat dan dinamis secara khusus di Indonesia, dengan pengguna secara aktif mendiskusikan isu-isu sosial, politik, dan budaya populer. Basis pengguna X Indonesia menunjukkan tingkat keterlibatan yang tinggi, sehingga menjadikannya sangat signifikan untuk mempelajari pola perilaku daring dan fenomena sosial.

Kemudahan akses dan kebebasan berekspresi yang ditawarkan oleh media sosial X juga menghasilkan dampak negatif berupa fenomena perundungan siber. Perundungan siber merujuk pada tindakan agresif dan disengaja yang dilakukan oleh individu atau kelompok menggunakan media elektronik, yang dapat terjadi kapan saja dan di mana saja, melampaui batasan fisik lingkungan tradisional. Studi terkini mengindikasikan bahwa perundungan siber memengaruhi sekitar 36% pengguna media sosial Indonesia, dengan korban mengalami konsekuensi psikologis yang

parah termasuk gangguan kecemasan, depresi, dan isolasi sosial (El Koshiry et al., 2024). Karakteristik media sosial X seperti potensi anonimitas, kemampuan diseminasi konten viral, dan kesingkatan pesan (keterbatasan karakter) menciptakan lingkungan yang sangat subur untuk perkembangan perilaku verbal yang toksik dan intimidatif. Perundungan siber di X dapat menjangkau audiens yang masif secara instan dan meninggalkan jejak digital permanen, berbeda dengan perundungan tatap muka, yang memperkuat efek berbahayanya. Kondisi ini menjadikan perundungan siber di X sebagai masalah sosial mendesak yang memerlukan solusi teknologi yang efektif.

Kecerdasan Buatan (Artificial Intelligence/AI), khususnya Pemrosesan Bahasa Alami (Natural Language Processing/NLP), menawarkan kerangka solusi yang menjanjikan untuk mengatasi masalah perundungan siber. NLP berfokus pada pengembangan sistem komputasi yang mampu secara otomatis memahami, menginterpretasi, dan memproses bahasa manusia. Teknik klasifikasi teks dapat dimanfaatkan dalam konteks perundungan siber untuk membangun model yang mampu secara otomatis mengidentifikasi dan memfilter konten yang mengandung elemen perundungan. Tinjauan sistematis komprehensif pada tahun 2023 mengonfirmasi bahwa penelitian terkini secara konsisten menunjukkan keunggulan model berbasis deep learning dibandingkan algoritma machine learning konvensional dalam menangani data yang besar dan kompleks, serta kemampuannya mengekstrak fitur secara otomatis tanpa rekayasa fitur manual yang ekstensif (Hasan et al., 2023).

Berbagai studi empiris telah membuktikan keunggulan model deep learning dalam konteks deteksi perundungan siber berbahasa Indonesia. Penelitian yang menerapkan arsitektur deep learning sekuensial menunjukkan bahwa model Bidirectional Long Short-Term Memory (BiLSTM) berhasil mencapai akurasi tinggi sebesar 81,82% pada data uji untuk klasifikasi komentar perundungan siber berbahasa Indonesia di platform Instagram (Iryani et al., 2025). Temuan ini menggarisbawahi kekuatan model berbasis LSTM dalam memahami urutan kalimat dan konteks, yang merupakan elemen krusial dalam mengidentifikasi pola perilaku perundungan yang tertanam dalam teks percakapan. Penelitian lain menunjukkan kinerja tinggi yang konsisten dari model Bi-LSTM dengan mencapai akurasi hingga 97% dalam mendeteksi berbagai jenis pelanggaran UU ITE, termasuk perundungan siber, di platform media sosial berbahasa Indonesia (Maulana & Aditya, 2025). Hasil-hasil ini memvalidasi efektivitas jaringan saraf rekuren dalam menangkap sifat sekuensial bahasa alami.

Model tunggal menghadapi keterbatasan inheren dalam menangkap fitur lokal dan pola sekuensial global secara bersamaan, meskipun arsitektur deep learning individual telah menunjukkan hasil yang menjanjikan. Efektivitas arsitektur CNN dalam klasifikasi teks berbahasa Indonesia telah divalidasi dalam berbagai konteks analisis bahasa natural. Penelitian menunjukkan bahwa arsitektur CNN dengan embedding layer, convolutional layer, pooling, dan dense layer dapat mencapai akurasi 99,10% dalam analisis sentimen komentar YouTube berbahasa Indonesia, dengan penekanan pada pentingnya preprocessing komprehensif yang mencakup text cleaning, normalization, tokenization, stopword removal, dan stemming (Cahyo et al., 2025). Dalam domain identifikasi konten problematik lainnya, CNN juga menunjukkan performa tinggi untuk identifikasi berita hoaks berbahasa Indonesia dari Facebook, mencapai akurasi 92,53% pada data pelatihan dan 81,09% pada data pengujian dengan pendekatan preprocessing serupa serta pemanfaatan Word2Vec untuk representasi vektor kata (Ramadhan et al., 2024). Penggunaan pre-trained word embeddings seperti Word2Vec terbukti meningkatkan pemahaman konteks semantik, aspek yang relevan untuk membedakan antara komentar bullying dan non-bullying dalam media sosial yang sering mengandung bahasa informal, slang, dan pencampuran kode.

Kesadaran akan keterbatasan model tunggal telah mendorong tren penelitian menuju arsitektur hibrida yang lebih canggih, yang kini dianggap sebagai salah satu pendekatan paling mutakhir dalam klasifikasi teks. Arsitektur hibrida Convolutional Neural Network - Long Short-Term Memory (CNN-LSTM) secara khusus menunjukkan potensi luar biasa karena kemampuannya menggabungkan kekuatan kedua model secara sinergis. CNN unggul dalam mengekstrak fitur lokal yang signifikan seperti frasa kunci, pola n-gram, atau kombinasi kata-kata toksik, sementara LSTM unggul dalam memahami konteks sekuensial, dependensi jangka panjang antar kata, dan alur sentimen keseluruhan dalam sebuah kalimat. Beberapa studi terkini telah memvalidasi efektivitas pendekatan hibrida ini pada data berbahasa Indonesia. Penelitian pada tahun 2023 berhasil menerapkan model hibrida CNN-LSTM pada data berbahasa Indonesia dari media sosial X dan mencapai akurasi 79,48% (Asqolani & Setiawan, 2023). Pada tahun yang sama, penelitian lain memperkuat temuan ini dengan menggunakan arsitektur hibrida CNN-BiLSTM yang ditingkatkan dengan ekspansi fitur FastText dan berhasil mencapai akurasi 80,55% pada data tweet Indonesia (Nasution & Setiawan, 2023). Konsistensi performa tinggi CNN dalam ekstraksi fitur lokal, dikombinasikan dengan kemampuan LSTM dalam memahami konteks sekuensial, menjadi fondasi kuat untuk pengembangan arsitektur hibrida CNN-LSTM dalam penelitian ini.

Beberapa gap penelitian masih tersisa dan perlu diatasi, meskipun model hibrida CNN-LSTM telah menunjukkan hasil yang menjanjikan. Pertama, penelitian terdahulu yang menerapkan arsitektur hibrida CNN-LSTM sering kali terbatas pada platform atau konteks data spesifik dan kurang memiliki evaluasi komprehensif khusus pada media sosial X, di mana pola linguistik dan manifestasi perundungan siber yang unik terjadi (Hafiza & Setiawan, 2025). Kedua, implementasi yang ada sering kali belum menginvestigasi secara sistematis dampak teknik regularisasi (seperti Early Stopping) terhadap generalisasi model, meskipun overfitting merupakan tantangan yang dikenal luas dalam deep

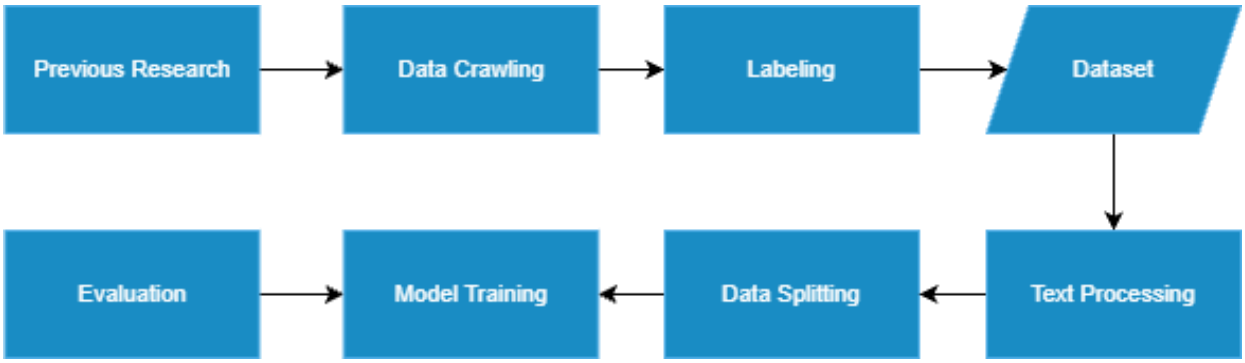
learning. Ketiga, peran embedding kata yang telah dilatih sebelumnya (seperti FastText) dalam meningkatkan kinerja model hibrida untuk bahasa Indonesia belum diteliti secara menyeluruh. Keempat, studi komparatif komprehensif yang memposisikan hibrida CNN-LSTM terhadap metode machine learning maupun model berbasis Transformer mutakhir masih langka. Penelitian menunjukkan bahwa kinerja model hibrida CNN-LSTM sangat dipengaruhi oleh proses optimasi hiperparameter dan pilihan desain arsitektural, yang tanpanya model mungkin tidak mencapai kinerja puncaknya (Faritha Banu et al., 2022). Kondisi ini mengindikasikan kebutuhan untuk mengembangkan model dasar yang kuat dengan kinerja yang diukur secara sistematis dan metodologi implementasi yang robust khusus untuk deteksi perundungan siber berbahasa Indonesia di media sosial X.

Penelitian ini bertujuan untuk mengisi gap tersebut melalui implementasi sistematis dan evaluasi yang ketat terhadap arsitektur hibrida CNN-LSTM. Penelitian ini secara spesifik berupaya menjawab pertanyaan-pertanyaan penelitian berikut: (1) Seberapa efektif Early Stopping dalam mencegah overfitting pada model CNN-LSTM untuk deteksi perundungan siber berbahasa Indonesia? (2) Sejauh mana embedding FastText yang telah dilatih sebelumnya meningkatkan kinerja klasifikasi model? (3) Bagaimana perbandingan model hibrida CNN-LSTM terhadap metode machine learning, arsitektur deep learning tunggal, dan model Transformer mutakhir dalam hal akurasi dan efisiensi komputasi? (4) Apa keseimbangan optimal antara kinerja model dan kompleksitas komputasi untuk deployment praktis?.

Penelitian ini memiliki empat kontribusi utama. Pertama, penelitian ini menyediakan implementasi dan evaluasi sistematis terhadap arsitektur hibrida CNN-LSTM yang dioptimalkan secara khusus untuk deteksi perundungan siber berbahasa Indonesia di media sosial X, dengan desain eksperimental komprehensif yang mencakup beragam skenario pengujian dan studi ablasi terkontrol. Kedua, penelitian ini menyajikan validasi empiris terhadap pentingnya kritis mekanisme Early Stopping dalam mencegah overfitting, dengan mengkuantifikasi dampaknya melalui eksperimen komparatif terkontrol yang mengungkapkan degradasi akurasi hingga 32% tanpa teknik regularisasi ini. Ketiga, penelitian ini melakukan studi komparatif komprehensif lintas paradigma arsitektural yang berbeda, termasuk arsitektur hibrida (CNN-LSTM), model deep learning tunggal (CNN dan LSTM standalone), metode machine learning (SVM, Naive Bayes), dan model berbasis Transformer mutakhir (IndoBERTweet), untuk menyediakan tolok ukur kinerja yang jelas bagi para peneliti. Keempat, penelitian ini mengembangkan model dasar yang efisien dan praktis yang mencapai keseimbangan yang menguntungkan antara kinerja deteksi (akurasi 88,38%) dan efisiensi komputasi (4,97 juta parameter), sehingga menjadikannya sangat cocok untuk deployment di lingkungan dengan sumber daya terbatas.

II. Studi Pustaka

Bagian ini menjelaskan metode dan tahapan penelitian yang digunakan secara detail. Penjelasan mencakup tahapan mulai dari studi literatur, pengumpulan dan persiapan data, perancangan dan optimasi model klasifikasi menggunakan arsitektur hibrida CNN-LSTM, hingga evaluasi kinerja model. Alur penelitian secara keseluruhan diilustrasikan pada Gambar 1.



Gambar 1. Tahapan Penelitian

A. Tahapan Penelitian

Penelitian ini dilakukan melalui serangkaian tahapan yang sistematis dan terstruktur. Dimulai dari studi literatur untuk memahami dasar teoretis dan penelitian terkait, diikuti dengan pengumpulan dan persiapan data mentah dari media sosial X. Data kemudian melalui tahap pelabelan manual dan preprocessing untuk membersihkan dan mengubahnya ke dalam format yang sesuai. Selanjutnya, data dibagi menjadi dataset pelatihan, validasi, dan pengujian. Model dasar dengan arsitektur hibrida CNN-LSTM kemudian dirancang dan dilatih menggunakan data

pelatihan dan divalidasi. Sebagai tahap penelitian lanjutan, optimasi hiperparameter sistematis dilakukan untuk menemukan konfigurasi model terbaik. Tahap akhir adalah pengujian dan evaluasi model akhir menggunakan data uji untuk mengukur kinerja secara komprehensif dan menarik kesimpulan.

Studi literatur komprehensif dilakukan pada tahap awal. Fokus meliputi deteksi perundungan siber pada teks berbahasa Indonesia, khususnya dari media sosial X, dan penerapan arsitektur Natural Language Processing (NLP) mulai dari machine learning konvensional hingga deep learning termasuk LSTM, CNN, dan model hibrida. Tinjauan literatur ini membangun fondasi teoretis yang kuat, mengidentifikasi kesenjangan penelitian, dan memvalidasi pemilihan metode. Sebagai hasilnya, arsitektur hibrida CNN-LSTM diadopsi. Pilihan ini didasarkan pada kemampuannya untuk menyeimbangkan ekstraksi fitur melalui CNN dengan pemahaman konteks sekuensial melalui LSTM, serta adanya kesenjangan optimasi untuk data teks informal di Indonesia.

B. Pengumpulan Data

Sumber data untuk penelitian ini adalah komentar teks berbahasa Indonesia yang berasal dari media sosial X. Akuisisi data dilakukan melalui Application Programming Interface (API) yang disediakan oleh platform, dengan memfilter komentar berdasarkan kata kunci dan hashtag yang relevan dengan topik. Untuk memastikan kualitas dan orisinalitas dataset, beberapa kriteria ditetapkan, seperti periode pengumpulan data dan memastikan hanya tweet asli, bukan retweet, yang dimasukkan dalam pengumpulan data. Contoh hasil dataset dapat dilihat pada Tabel 1.

Pelabelan manual dilakukan setelah pengumpulan data. Tahap ini sangat penting dalam supervised learning karena model memerlukan ground truth sebagai kunci jawaban untuk pembelajaran. Setiap unit data (tweet) dianalisis dan diklasifikasikan secara individual sebagai 'bullying' (nilai 1) atau 'non-bullying' (nilai 0). Objektivitas dan konsistensi pelabelan dijaga dengan merujuk pada pedoman anotasi yang telah disusun. Dataset akhir sebanyak 13.677 komentar diperoleh setelah pengumpulan dan pelabelan data. Distribusi label dalam dataset ini menunjukkan komposisi data. Contoh dataset yang telah dilabeli dapat dilihat pada Tabel 2.

C. Preprocessing Data

Data yang telah dilabeli memerlukan *preprocessing* sebelum pelatihan model *deep learning*. Tahap ini bertujuan untuk membersihkan, menstandarkan, dan mengonversi teks ke dalam format numerik terstruktur. Tahapan selanjutnya meliputi *cleaning*, *case folding*, *tokenization*, dan *padding*. *Text cleaning* adalah tahap *preprocessing* awal untuk seluruh *dataset*. Proses ini menghilangkan noise seperti URL, *mention* (@username), *hashtag*, dan karakter khusus yang mengganggu pembelajaran model. Semua huruf dikonversi menjadi huruf kecil melalui *case folding*. Langkah ini memastikan standardisasi dan konsistensi teks sebelum pemrosesan selanjutnya. Setiap kalimat dalam kolom *tweet* yang telah dibersihkan dipecah menjadi unit kata (*token*) melalui proses *tokenization*. Kelas *Keras Tokenizer* digunakan untuk membangun kosakata hanya dari data pelatihan untuk mencegah kebocoran data, dibatasi hingga 15.000 kata unik yang paling sering muncul. Model CNN-LSTM memerlukan input yang seragam, sehingga *padding* dilakukan pada komentar sekuensial dengan panjang yang berbeda. Setiap sekuens distandarkan menjadi 100 *token* di mana sekuens pendek ditambahkan dengan nilai nol (*post-padding*), sedangkan sekuens panjang dipotong di akhir (*post-truncation*).

D. Pembagian Data

Dataset akhir yang telah melalui *preprocessing* dibagi menjadi tiga bagian independen untuk keperluan pemodelan: 8.752 data pelatihan, 2.189 data validasi, dan 2.736 data uji. Pembagian ini setara dengan rasio 80:10:10 yang dirancang agar proporsi data pelatihan cukup besar, tetapi tetap menyisakan data yang memadai untuk validasi dan pengujian objektif. Strategi pembagian ini memastikan bahwa model dilatih pada data yang cukup sambil mempertahankan *set* validasi dan uji yang independen untuk evaluasi kinerja yang tidak bias.

E. Arsitektur Model Yang Diusulkan

Tahap pelatihan model adalah inti dari proses eksperimental, di mana arsitektur hibrida CNN-LSTM dilatih dari awal (*from scratch*). Berbeda dengan *fine-tuning* yang mengadaptasi model yang telah dilatih sebelumnya, pendekatan ini melatih model mulai dengan pengetahuan kosong menggunakan bobot acak untuk mempelajari pola secara eksklusif dari *dataset* yang telah disiapkan, seperti yang diilustrasikan pada Gambar 2.

Tabel 1. Sampel Pengumpulan Data

Pengguna	Teks
User1	Goblok ini Hoaxs ..Tolong Kalo Jadi Pendukung Anies terus Jangan Gilak Bikan Hoaxs .. Tolong @DivHumas_Polri ini ditindaklanjuti .
User2	Dialog Lingkungan Hidup Ustadz Abdul Somad dan Rocky Gerung di Tanjung Belit https://t.co/NLTPgeBuka

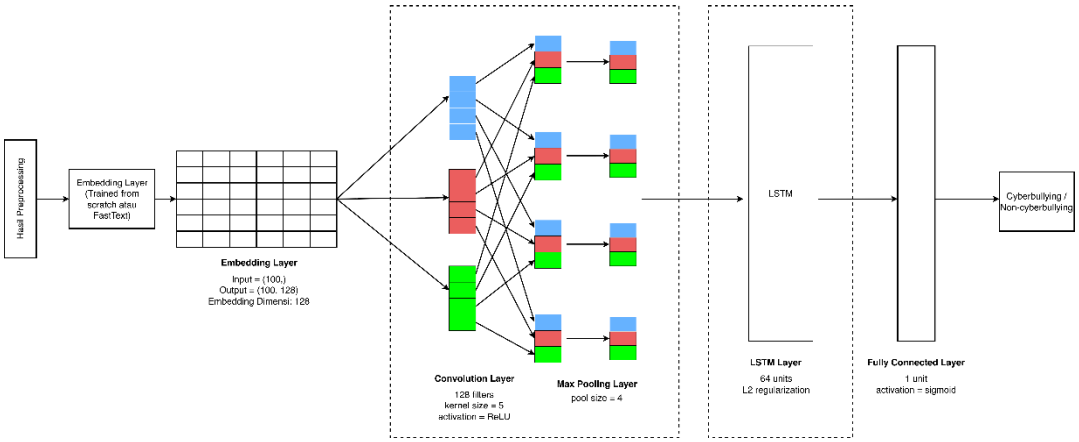
User3	Gibran Datang ke Rumahnya Rocky Gerung: Saya Kasih Kopi Oke You Bicara Anak Muda! https://t.co/ewhKdZIGGa

Tabel 2. Data Sampel Yang Diberi Label

Pengguna	Teks	Label
User1	Goblok ini Hoaxs ..Tolong Kalo Jadi Pendukung Anies terus Jangan Gilak Bikan Hoaxs .. Tolong @DivHumas_Polri ini ditindaklanjuti .	1
User2	Dialog Lingkungan Hidup Ustadz Abdul Somad dan Rocky Gerung di Tanjung Belit https://t.co/NLTPgeBuka	0
User3	Gibran Datang ke Rumahnya Rocky Gerung: Saya Kasih Kopi Oke You Bicara Anak Muda! https://t.co/ewhKdZIGGa	0

Tabel 3. Konfigurasi Model Dasar CNN-LSTM

Pengguna	Teks
Embedding Layer	Trainable, dimensi vektor kata = 128, dilatih dari nol
Conv1D Layer	128 filter, ukuran kernel = 5
BatchNormalization	Digunakan untuk stabilitas pembelajaran
Activation	ReLU



Gambar 2. CNN-LSTM

Tahap Arsitektur dasar yang diusulkan terdiri dari beberapa lapisan yang diatur secara sekuensial. Lapisan pertama adalah *embedding layer* dengan ukuran kosakata 15.000 kata, dimensi *embedding* 128, dan panjang sekuens maksimum 100 *token*. Lapisan ini mengonversi indeks kata menjadi representasi vektor padat. Setelah *embedding layer*, lapisan konvolusional satu dimensi (*Conv1D*) dengan 128 *filter* dan ukuran *kernel* 5 diterapkan menggunakan fungsi aktivasi ReLU dan *same padding* untuk mempertahankan panjang sekuens. Lapisan ini mengekstrak fitur lokal dari sekuens yang telah di-*embed*. Lapisan *max pooling* dengan ukuran *pool* 2 kemudian mengurangi dimensi spasial menjadi setengah, mempertahankan fitur yang paling menonjol. Fitur yang telah di-*pool* dimasukkan ke dalam lapisan LSTM dengan 64 unit, yang menggabungkan *dropout* 0,2 dan *recurrent dropout* 0,2 untuk mencegah *overfitting*. Lapisan LSTM menangkap dependensi jangka panjang dan pola sekuensial dalam teks. Akhirnya, lapisan *dense output* dengan satu unit dan fungsi aktivasi *sigmoid* menghasilkan probabilitas klasifikasi biner. Model dikompilasi menggunakan *optimizer* Adam dengan *learning rate* 0,001 dan fungsi *loss binary crossentropy*. Untuk mencegah *overfitting*, *callback Early Stopping* diimplementasikan untuk memantau *validation loss* dengan *patience* 3 *epoch*, secara otomatis mengembalikan bobot model terbaik ketika kinerja validasi berhenti meningkat. Konfigurasi hiperparameter terperinci yang digunakan disajikan pada Tabel 4.

F. Skenario Eksperimental

Untuk mengevaluasi dan menemukan arsitektur yang paling efektif, penelitian ini merancang beberapa skenario pengujian sistematis. Model dasar yang diusulkan adalah arsitektur hibrida CNN-LSTM dengan konfigurasi yang ditunjukkan pada Tabel 3. Untuk menguji dan memvalidasi kinerja model dasar ini, serta untuk mengeksplorasi potensi peningkatannya, beberapa skenario pengujian sistematis dirancang sebagai berikut.

Skenario pertama berfokus pada evaluasi model dasar, di mana model utama dilatih dan dievaluasi menggunakan callback EarlyStopping untuk menetapkan kinerja sebagai tolok ukur utama. Skenario kedua menganalisis peran

EarlyStopping melalui pendekatan komparatif di mana model dasar dilatih untuk epoch penuh tanpa EarlyStopping untuk membuktikan secara empiris dampak overfitting dan memvalidasi pentingnya penggunaan callback ini. Skenario ketiga melakukan studi komparatif lanjutan dengan serangkaian eksperimen tambahan untuk memperkaya analisis, termasuk implementasi pre-trained FastText embeddings dibandingkan dengan embedding yang dilatih dari awal, serta perbandingan dengan model lain yaitu IndoBERTweet untuk mengevaluasi sejauh mana model transformer khusus bahasa Indonesia yang dilatih pada data media sosial X dapat mengungguli atau melengkapi arsitektur CNN-LSTM dalam tugas deteksi perundungan siber. Selain itu, optimasi hiperparameter menggunakan Keras Tuner dengan strategi BayesianOptimization juga dilakukan untuk menemukan konfigurasi terbaik dan memaksimalkan kinerja model yang diuji.

G. Evaluasi Model

Evaluasi kinerja model adalah tahap akhir penelitian. Evaluasi ini secara objektif menguji kemampuan generalisasi model menggunakan *dataset* uji. Kinerja model dianalisis baik secara kuantitatif maupun kualitatif. Analisis kuantitatif dimulai dengan membuat *Confusion Matrix* untuk memvisualisasikan hasil prediksi. Dari matriks ini, metrik evaluasi standar untuk tugas klasifikasi dihitung. *Accuracy* mengukur persentase prediksi yang benar dari semua data, dihitung sebagai rasio *true positive* dan *true negative* terhadap jumlah total sampel. *Precision* mengukur akurasi prediksi positif, merepresentasikan proporsi kasus positif yang diprediksi dengan benar di antara semua kasus yang diprediksi positif. *Recall* mengukur kemampuan model untuk menemukan semua instans positif, merepresentasikan proporsi kasus positif yang diidentifikasi dengan benar di antara semua kasus positif aktual. *F1-Score* merepresentasikan rata-rata harmonik dari *Precision* dan *Recall*, memberikan ukuran seimbang yang mempertimbangkan baik *false positive* maupun *false negative*. *Area Under the Curve* (AUC) mengukur kemampuan model untuk membedakan antara kelas *cyberbullying* dan *non-cyberbullying* di berbagai *threshold* klasifikasi, dengan nilai yang lebih dekat ke 1,0 menunjukkan kinerja diskriminatif yang lebih baik. Selain itu, analisis kualitatif juga dilakukan dengan meninjau beberapa contoh komentar yang salah diklasifikasikan oleh model melalui analisis kesalahan untuk memahami karakteristik teks di mana model mengalami kesulitan. Hasil dari metrik ini akan menjadi dasar utama untuk menganalisis dan menyimpulkan tingkat keberhasilan dan keandalan model hibrida CNN-LSTM yang telah dikembangkan.

Tabel 4. Konfigurasi Hyperparameter Model CNN-LSTM

Pengguna	Teks
Ukuran Vocabulary	15.000 kata
Panjang Sequence	100 token
Dimensi Embedding (Baseline)	128
Dimensi Embedding (FastText)	300
Conv1D Filters	128
Conv1D Kernel Size	5
LSTM Units	64
Dropout Rate	0.5
Batch Size	32
Learning Rate	0.001
Optimizer	Adam
Embedding Layer	Trainable, dimensi vektor kata = 128, dilatih dari nol
Conv1D Layer	128 filter, ukuran kernel = 5
BatchNormalization	Digunakan untuk stabilitas pembelajaran
Activation	ReLU

III. HASIL DAN PEMBAHASAN

Penelitian ini menyajikan evaluasi komprehensif terhadap kinerja arsitektur *hybrid* CNN-LSTM yang diusulkan untuk mendeteksi ucapan *cyberbullying* pada dataset media sosial X berbahasa Indonesia. Serangkaian eksperimen dilakukan dalam lingkungan komputasi berkemampuan tinggi menggunakan Google Colaboratory dengan akselerasi GPU Tesla T4, memanfaatkan *framework* TensorFlow dan Keras. Dataset yang digunakan berjumlah 13.677 titik data, dibagi secara proporsional dengan rasio 80:10:10 menjadi data pelatihan (8.752 sampel), data validasi (2.189 sampel), dan data uji (2.736 sampel). Evaluasi dilakukan berdasarkan tiga skenario pengujian sistematis yang dirancang dalam

metodologi: (1) evaluasi kinerja model dasar, (2) analisis strategi pelatihan dan dampak regularisasi, dan (3) studi arsitektur perbandingan dan pembandingan dengan metode terkini.

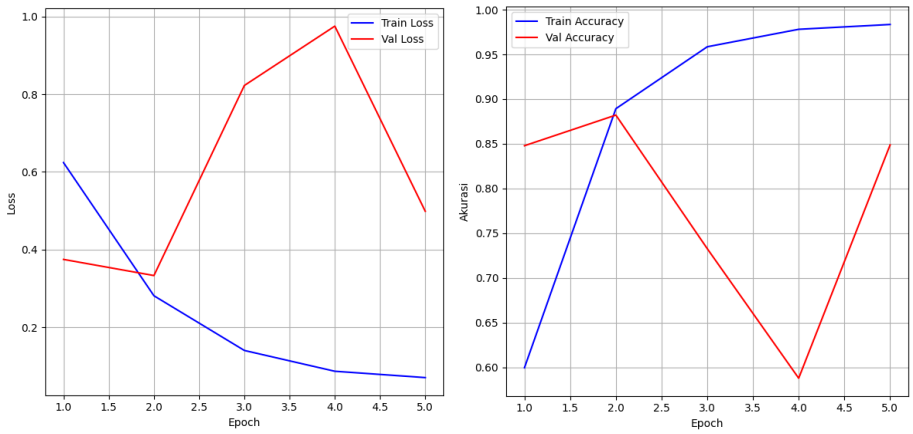
A. Kinerja Model Dasar

Skenario pertama bertujuan untuk menetapkan kinerja dasar menggunakan model CNN-LSTM yang dilatih dengan representasi kata yang diinisialisasi secara acak dari awal dengan *embedding* 128 dimensi. Model ini dilengkapi dengan mekanisme *Early Stopping* untuk secara otomatis menghentikan pelatihan ketika tidak ada peningkatan kinerja pada data validasi, guna memperoleh bobot model optimal dan mencegah *overfitting*. Konfigurasi hiperparameter yang digunakan secara rinci disajikan dalam Tabel 4.

Stabilitas proses pelatihan dievaluasi melalui visualisasi kurva pembelajaran. Gambar 3a menampilkan grafik *Loss* pelatihan dan validasi, sementara Gambar 3b menampilkan grafik Akurasi. Kedua grafik menunjukkan konvergensi model yang efektif, di mana mekanisme *Early Stopping* berhasil menghentikan pelatihan pada *epoch* ke-5 tepat sebelum *Loss* validasi mulai menunjukkan tren kenaikan, sehingga mencegah model dari *overfitting*.

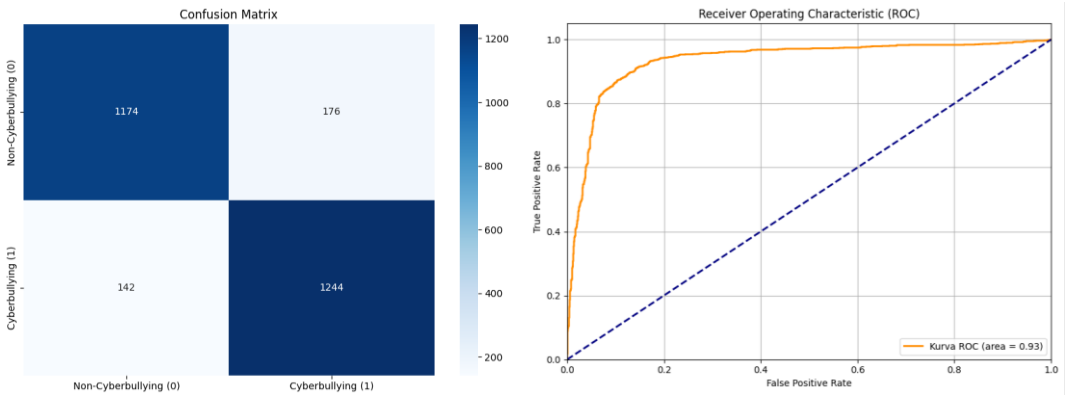
Tabel 5. Rincian Hasil Evaluasi Model CNN-LSTM

Metrik	Nilai
Akurasi	0.8838
Presisi	0.88
Recall	0.88
F1 Score	0.88
AUC	0.9346



(a) (b)

Gambar 3. Kurva *Loss* (a) dan Akurasi (b) latih validasi pada Model *CNN-LSTM* dengan *Early*



(a) (b)

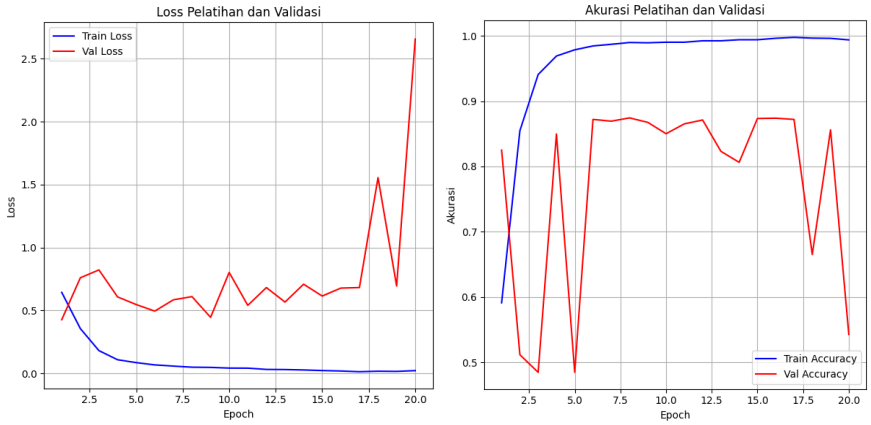
Gambar 4. *Confusion Matrix* data uji (c) dan Kurva ROC data latih-validasi (d) pada Model *Baseline* CNN-LSTM dengan *Early Stopping*.

Berdasarkan hasil uji pada *set* uji yang disisihkan, model dasar menunjukkan kinerja yang sangat solid. Hasil evaluasi kuantitatif dirangkum dalam Tabel 5. Model berhasil mencapai tingkat akurasi 88,38% dengan skor *F1* sebesar 88,00%. Untuk memahami distribusi kesalahan prediksi model secara lebih mendalam, analisis dilakukan menggunakan *Confusion Matrix* yang ditampilkan pada Gambar 4a. Visualisasi ini menunjukkan bahwa model memiliki kemampuan seimbang dalam mengenali kelas mayoritas dan minoritas. Secara spesifik, tingkat *Recall* untuk kelas positif (*cyberbullying*) mencapai 90%. Tingkat *Recall* yang tinggi ini sangat krusial dalam sistem deteksi konten berbahaya untuk meminimalkan risiko komentar *bullying* yang terlewat (*false negatives*).

Selain itu, kemampuan diskriminatif model diverifikasi melalui kurva ROC yang ditampilkan pada Gambar 4b. Nilai *Area Under Curve* (AUC) yang tinggi sebesar 0,9346 menunjukkan bahwa arsitektur CNN-LSTM memiliki keandalan yang sangat baik dalam memisahkan probabilitas antara kelas *cyberbullying* dan *non-cyberbullying* pada berbagai ambang batas, mengonfirmasi posisinya sebagai *baseline* yang kuat untuk penelitian ini.

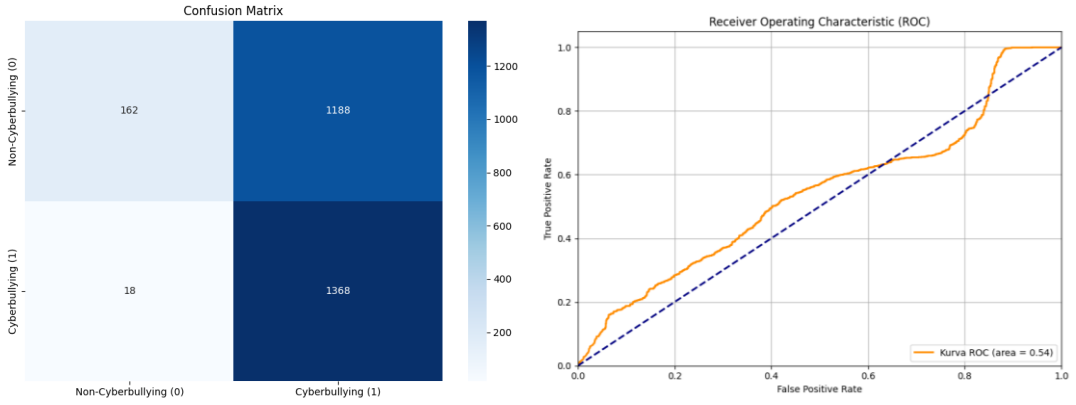
Tabel 6. Perbandingan Model Dengan Dan Tanpa Early Stopping

Model	Akurasi	Presisi	Recall	F1 Score	AUC
CNN-LSTM (Baseline)	0.8838	0.88	0.88	0.88	0.9346
CNN-LSTM (Tanpa Early Stopping)	0.5592	0.72	0.56	0.46	0.5266



(a) (b)

Gambar 5. Kurva *Loss* (a) dan Akurasi (b) latih validasi pada Model *CNN-LSTM* tanpa *Early Stopping*.



(a) (b)

Gambar 6. *Confusion Matrix* data uji (c) dan Kurva ROC data latih-validasi (d) pada Model *Baseline* CNN-LSTM tanpa *Early Stopping*.

B. Pengaruh Strategi Pelatihan (Early Stopping)

Untuk memvalidasi secara empiris pentingnya strategi regulasi waktu dalam pelatihan Deep Learning, dilakukan uji perbandingan antara model yang dilatih dengan mekanisme Early Stopping (Skenario 1) dan tanpa mekanisme tersebut (Skenario 2). Hasil eksperimen menunjukkan dampak yang sangat signifikan dari strategi ini terhadap stabilitas model dan generalisasi.

Ketika pelatihan model diizinkan berjalan penuh selama 20 epoch tanpa mekanisme penghentian otomatis, fenomena overfitting yang parah terjadi. Seperti yang terlihat pada Gambar 5a (Loss) dan Gambar 5b (Accuracy), divergensi ekstrem (pemisahan) antara kinerja pelatihan dan validasi mulai terlihat mulai epoch ke-7. Kurva Loss validasi pada Gambar 5a meningkat tajam menjauh dari Loss pelatihan, menunjukkan bahwa model mulai menghafal data pelatihan (memorisasi) dan kehilangan kemampuan generalisasi pada data baru.

Tabel 7. Hasil Optimasi Hyperparameter

Hyperparameter	Nilai Baseline	Nilai Optimized
Conv1D Filters	128	256
LSTM Units	64	96
Dropout Rate	0.5	0.5
Learning Rate	0.001	0.0001

Dampak *overfitting* ini terlihat jelas pada penurunan kinerja pengujian. Meskipun akurasi pada data pelatihan terus meningkat hingga hampir sempurna (99,38%), kinerja model pada data pengujian mengalami penurunan drastis. Akurasi anjlok menjadi 55,92%, dan nilai AUC turun menjadi 0,5266, mendekati kinerja tebakan acak. Kegagalan model dalam melakukan klasifikasi yang benar akibat *overfitting* ditunjukkan melalui *Confusion Matrix* pada Gambar 6a dan kurva ROC yang hampir datar pada Gambar 6b.

Sebaliknya, dengan implementasi *Early Stopping* (Model Dasar), pelatihan dihentikan secara otomatis pada titik optimal (sekitar *epoch* 5), dan bobot model dipulihkan ke kondisi terbaiknya. Strategi ini terbukti berhasil dalam mempertahankan kemampuan generalisasi model dan menghasilkan model yang tangguh. Perbandingan kinerja kuantitatif antara dua kondisi pelatihan ini dirangkum dalam Tabel 6.

C. Pengaruh Embedding Yang Sudah Dilatih Dan Optimasi

Untuk mengevaluasi kontribusi berbagai komponen arsitektur terhadap performa deteksi cyberbullying, penelitian ini Penelitian lebih lanjut dilakukan dalam Skenario 3 untuk meningkatkan kinerja model dasar melalui penggunaan representasi kata yang telah dilatih sebelumnya (*pre-trained FastText embeddings*) dan optimasi hiperparameter otomatis.

Secara empiris, penggunaan *FastText Embeddings* (dimensi 300) terbukti memberikan peningkatan dalam stabilitas konvergensi dan kemampuan diskriminasi model dibandingkan dengan inisialisasi acak. Nilai AUC meningkat dari 0,9346 menjadi 0,9559, dan akurasi sedikit meningkat menjadi 88,49%. Peningkatan ini menunjukkan bahwa representasi vektor kata yang dilatih pada korpus Wikipedia Indonesia yang besar membantu model memahami konteks semantik kata-kata langka atau *slang* yang sering ditemukan di media sosial, yang mungkin tidak terwakili dengan baik jika hanya mengandalkan data pelatihan yang terbatas.

Di sisi lain, upaya optimasi hiperparameter menggunakan metode *Bayesian Optimization* (*Keras Tuner*) menghasilkan konfigurasi arsitektur dengan kapasitas lebih besar, yaitu 256 *filter* pada lapisan CNN dan 96 unit pada lapisan LSTM, tetapi dengan *learning rate* yang lebih konservatif (0,0001). Meskipun model yang dioptimalkan ini menghasilkan AUC yang sangat kompetitif (0,9507), peningkatan akurasi dan *F1-Score* relatif marginal dibandingkan dengan *baseline*. Hal ini menunjukkan bahwa konfigurasi *baseline* yang diusulkan sudah cukup *robust* dan optimal untuk karakteristik dataset ini, dan penambahan kompleksitas model tidak selalu berbanding lurus dengan peningkatan akurasi yang signifikan.

Dinamika pelatihan, yang divisualisasikan melalui kurva akurasi dan *loss*, menunjukkan bahwa model mulai menunjukkan tanda-tanda *overfitting* setelah *epoch* ke-2, ditandai dengan peningkatan *loss* validasi dan penurunan akurasi validasi. Namun, mekanisme *Early Stopping* secara efektif intervensi dengan menghentikan pelatihan pada *epoch* 5 dan mengembalikan bobot model ke kondisi optimalnya pada *epoch* 2. Nilai AUC validasi terbaik sebesar 0,9289 dicapai pada *checkpoint* ini, mengonfirmasi bahwa strategi *Early Stopping* berhasil mempertahankan kemampuan generalisasi model. Hasil optimasi hiperparameter disajikan dalam Tabel 7.

Proses optimasi mengidentifikasi konfigurasi dengan *filter* Conv1D yang digandakan (256 vs. 128) dan unit LSTM yang ditingkatkan (96 vs. 64), serta *learning rate* yang dikurangi (0,0001 vs. 0,001). Menariknya, model yang dioptimalkan mencapai kinerja yang sebanding dengan model dasar dalam hal akurasi (88,05% vs. 88,38%) dan *F1-Score* (0,88 untuk keduanya), dengan penurunan akurasi yang sedikit (-0,33 poin persentase). Namun, serupa dengan eksperimen *FastText*, model yang dioptimalkan menunjukkan kinerja diskriminatif yang lebih unggul dengan AUC

0,9507, mewakili peningkatan relatif 1,61% dibandingkan dengan model dasar. Hal ini menunjukkan bahwa peningkatan kapasitas model (5,21 juta parameter vs. 4,69 juta) dan *learning rate* yang lebih konservatif memungkinkan kalibrasi probabilitas yang lebih baik di berbagai ambang klasifikasi, meskipun perkiraan akurasi titik tetap serupa. Skor *F1* yang sebanding menunjukkan bahwa konfigurasi *baseline* dan yang dioptimalkan mencapai *trade-off precision-recall* yang serupa, menunjukkan bahwa hiperparameter *baseline* sudah cukup sesuai untuk tugas ini. Peningkatan AUC yang marginal, meskipun secara statistik signifikan, datang dengan biaya kompleksitas komputasi dan waktu pelatihan yang lebih tinggi.

D. Studi Arsitektur Perbandingan (Ablation Study)

Untuk memvalidasi efektivitas desain hybrid yang menggabungkan CNN dan LSTM, dilakukan perbandingan komprehensif (ablation study) terhadap berbagai variasi arsitektur Deep Learning. Studi ini membandingkan model hybrid dengan model tunggal (CNN, LSTM, Bi-LSTM) serta variasi hybrid lainnya (CNN-BiGRU). Semua model dalam studi ini menggunakan FastText Embeddings untuk memastikan perbandingan yang adil.

Berdasarkan rekapitulasi hasil eksperimen, model tunggal Bi-LSTM dengan Attention mencatat kinerja tertinggi di antara semua varian Deep Learning klasik yang diuji, dengan pencapaian F1-Score 88,77% dan AUC 0,9577. Keunggulan model ini menunjukkan bahwa mekanisme attention dan pemahaman konteks bidirectional sangat efektif dalam menangkap fitur linguistik cyberbullying yang seringkali bergantung pada konteks kalimat yang lengkap.

Meskipun demikian, model hybrid CNN-BiGRU dan CNN-LSTM masih menunjukkan kinerja yang sangat kompetitif dengan perbedaan kinerja kurang dari 0,5% dibandingkan dengan model terbaik. Lebih jauh lagi, arsitektur hybrid ini menawarkan keunggulan dalam efisiensi komputasi. Lapisan konvolusional (CNN) di awal arsitektur berfungsi untuk mengurangi dimensi fitur secara efektif sebelum diproses oleh lapisan recurrent, sehingga mempercepat waktu pelatihan dan inference dibandingkan dengan model Bi-LSTM murni yang harus memproses urutan kata secara lengkap satu per satu.

Arsitektur CNN saja menunjukkan kinerja yang luar biasa dalam hal precision (90,71%), mengindikasikan kemampuan kuat model ini dalam mengidentifikasi instance positif dengan false positive minimal. Konvergensi cepat yang dicapai oleh CNN (berhenti pada epoch 6) mencerminkan efisiensi komputasinya dalam mengekstraksi fitur n-gram lokal. Namun, nilai recall yang sedikit lebih rendah (85,93%) menunjukkan keterbatasan model ini dalam menangkap konteks global kalimat yang lengkap dibandingkan dengan model berbasis sekuensial. Sebaliknya, arsitektur LSTM murni mencapai kinerja yang seimbang di semua metrik tetapi menuntut biaya komputasi yang signifikan, memerlukan 27 epoch untuk konvergen. Hal ini mengonfirmasi sifat iteratif dari pemrosesan sekuensial di LSTM yang lebih lambat dibandingkan dengan pemrosesan paralel di CNN.

Dalam kategori model hybrid, varian CNN-BiLSTM dan CNN-BiGRU menunjukkan kinerja yang sangat sebanding dengan akurasi sekitar 88%. Temuan menarik adalah kesetaraan kinerja antara BiLSTM dengan akurasi 0,8816 dan BiGRU dengan nilai akurasi 0,8834 dengan jumlah parameter yang identik. Hal ini mengindikasikan bahwa mekanisme gating yang lebih sederhana di GRU dapat menyamai efektivitas arsitektur LSTM yang lebih kompleks untuk tugas ini, tetapi dengan efisiensi operasional yang lebih baik. Model yang diusulkan, CNN-LSTM (Baseline + FastText), menempati posisi keseimbangan optimal antara kinerja (F1-Score 88%) dan efisiensi (konvergensi dalam 6 epoch) dengan ukuran model yang relatif kompak.

Temuan paling signifikan adalah munculnya arsitektur BiLSTM + Attention sebagai model dengan kinerja terbaik secara keseluruhan, mencapai F1-Score 88,77% dan AUC tertinggi sebesar 0,9577. Kemampuan mekanisme attention untuk secara dinamis memfokuskan bobot pada kata-kata penting (salient) terbukti memberikan keunggulan diskriminatif, terutama dalam meminimalkan kesalahan prediksi. Meskipun demikian, keunggulan ini datang dengan peningkatan kompleksitas model dan waktu pelatihan yang lebih lama. Mengingat perbedaan kinerja yang marginal di antara semua arsitektur canggih ini (rentang perbedaan akurasi hanya 0,66%), pemilihan model akhir untuk implementasi praktis harus mempertimbangkan faktor efisiensi komputasi dan kendala deployment, di mana arsitektur hybrid CNN-BiGRU atau CNN-LSTM menawarkan alternatif yang lebih ringan dibandingkan dengan model berbasis Attention yang berat. Hasil komprehensif dari ablation study ini dirangkum dalam Tabel 8.

Tabel 8. Rekapitulasi Kinerja Komprehensif Seluruh Model

Model	Akurasi	Presisi	Recall	F1-Score	AUC
IndoBERTweet	**0.915**	**0.92**	**0.92**	**0.915**	-
BiLSTM + Attention	0.8871	0.8945	0.8810	0.8877	0.9577
CNN-LSTM (+ FastText)	0.8849	0.89	0.88	0.88	0.9559
CNN	0.8841	0.9071	0.8593	0.8825	0.9566
CNN-BiGRU	0.8834	0.8834	0.8834	0.8834	0.9566
LSTM	0.88	0.88	0.88	0.88	0.9596
CNN-BiLSTM	0.8816	0.8820	0.8816	0.8815	0.9548

CNN-LSTM (Base Line Early Stopping)	0.8838	0.88	0.88	0.88	0.9346
CNN-LSTM (Keras Tuner)	0.8805	0.88	0.88	0.88	0.9507
SVM + TF-IDF	0.7818	0.8266	0.7818	0.7747	-
CNN-LSTM (Tanpa EarlyStopping)	0.5592	0.72	0.56	0.46	0.5266

E. Benchmarking dengan Model State-of-the-Art

Untuk menempatkan kinerja arsitektur CNN-LSTM yang diusulkan dalam konteks yang lebih luas, dilakukan analisis perbandingan mendalam terhadap berbagai pendekatan, mulai dari metode Machine Learning klasik, variasi arsitektur Deep Learning, hingga model transformer state-of-the-art. Evaluasi ini bertujuan untuk memvalidasi keunggulan relatif dari metode yang diusulkan dibandingkan dengan pendekatan yang telah ada sebelumnya.

Sebagai baseline benchmark untuk metode konvensional, Support Vector Machine (SVM) dengan fitur TF-IDF dipilih untuk mewakili pendekatan Machine Learning klasik. Hasil evaluasi menunjukkan kesenjangan kinerja yang substansial antara pendekatan klasik dan Deep Learning. Model CNN-LSTM yang diusulkan mencatat akurasi sebesar 88,38%, mewakili peningkatan signifikan sebesar 10,2 poin persentase dibandingkan dengan akurasi SVM yang hanya mencapai 78,18%. Lebih kritis lagi, peningkatan F1-Score sebesar 10,53 poin persentase pada model Deep Learning menunjukkan keseimbangan yang jauh lebih baik antara precision dan recall. Analisis mendalam terhadap kinerja SVM mengungkapkan pola asimetris, di mana meskipun model dapat mencapai precision tinggi untuk kelas cyberbullying, nilai recall-nya sangat rendah yaitu 60,32%. Hal ini mengindikasikan bahwa model klasik cenderung gagal dalam mendeteksi hampir setengah dari total komentar cyberbullying yang sebenarnya, sebuah kelemahan fatal dalam sistem deteksi yang memprioritaskan keselamatan pengguna. Sebaliknya, arsitektur CNN-LSTM menunjukkan kinerja yang seimbang pada kedua metrik dengan nilai 0,88, mengonfirmasi superioritas representasi fitur otomatis dalam menangkap nuansa bahasa yang kompleks.

Selanjutnya, model yang diusulkan dibandingkan dengan IndoBERTweet, model transformer berbasis BERT yang secara khusus dilatih pada korpus Twitter Indonesia skala masif, mewakili state-of-the-art saat ini. Hasil pengujian memposisikan IndoBERTweet sebagai model dengan kinerja tertinggi di semua metrik evaluasi, dengan pencapaian akurasi dan F1-Score sebesar 91,59%. Keunggulan ini dapat dikaitkan dengan proses pre-training yang ekstensif yang memberikan model pemahaman mendalam tentang pola linguistik, slang, dan struktur informal yang khas dari media sosial Indonesia. Namun, temuan yang patut dicatat adalah bahwa model Deep Learning terbaik yang dikembangkan dalam penelitian ini, yaitu BiLSTM dengan Attention, mencapai akurasi 88,71%, hanya tertinggal sekitar 2,88 poin persentase dari IndoBERTweet.

Kesenjangan kinerja yang relatif sempit ini menjadi sangat signifikan ketika mempertimbangkan disparitas sumber daya komputasi antara kedua pendekatan. Model IndoBERTweet memiliki kompleksitas parameter yang mencapai sekitar 110 juta, jauh melebihi model BiLSTM dengan Attention yang hanya memiliki sekitar 5,32 juta parameter. Selain itu, IndoBERTweet memerlukan 50 epoch pelatihan untuk mencapai konvergensi optimal, sementara model BiLSTM dan CNN-LSTM hanya memerlukan kurang dari 10 epoch. Hal ini menunjukkan bahwa arsitektur berbasis RNN dan CNN yang diusulkan menawarkan efisiensi komputasi yang jauh lebih tinggi, kecepatan inference yang lebih baik, dan kelayakan deployment pada perangkat dengan sumber daya terbatas, tanpa mengorbankan akurasi secara drastis.

Evaluasi komprehensif terhadap semua varian model yang diuji, seperti yang dirangkum dalam Tabel 10, menunjukkan bahwa arsitektur hybrid CNN-LSTM dan variannya secara konsisten menempati posisi kinerja yang kompetitif antara baseline klasik dan model transformer yang berat. Model CNN-BiGRU, misalnya, mencapai akurasi 88,34% dengan efisiensi tinggi, setara dengan model LSTM murni tetapi dengan struktur yang lebih kompak. Sementara itu, penggunaan FastText embeddings di CNN-LSTM terbukti meningkatkan kemampuan diskriminatif model, tercermin dalam peningkatan nilai AUC menjadi 0,9559. Temuan-temuan ini memvalidasi bahwa arsitektur hybrid fundamental yang menggabungkan ekstraksi fitur lokal dari CNN dengan pemodelan sekuensial dari LSTM adalah pendekatan yang robust dan efektif untuk karakteristik linguistik cyberbullying media sosial Indonesia, menawarkan keseimbangan optimal antara kinerja deteksi dan efisiensi sumber daya.

Tabel 9. Contoh Kesalahan Prediksi Model CNN-LSTM Pada Data Uji

No	Teks Komentar (Asli)	Label Asli	Prediksi Model	Status Kesalahan
1	masuk tan anjing	Non-Bullying (0)	Bullying (1)	False Positive

2	goodness parent actually watching something apple tv brilliant yet another streaming service	Non-Bullying (0)	Bullying (1)	False Positive
3	aduh lak onok rawan disalahgunakan men	Non-Bullying (0)	Bullying (1)	False Positive
4	jomok bawa barak iya oppa kdm iya anjr panggil sayang mau pacar panggil sayang first tonton mv kambek jeoang skz ya aju main vocalnya oek	Bullying (1)	Non-Bullying (0)	False Negative
5	halo non moot ijin ikut ya selamat lolo seleksi kerja tysm moga halus pity sylus nya bandel ya	Non-Bullying (0)	Bullying (1)	False Positive

F. Analisis Kesalahan Kualitatif

Sesuai dengan desain evaluasi untuk melengkapi metrik kuantitatif, dilakukan tinjauan mendalam terhadap sampel data uji yang salah diklasifikasikan oleh model yang diusulkan (CNN-LSTM). Analisis ini bertujuan untuk mengidentifikasi karakteristik linguistik spesifik dan pola teks yang menjadi tantangan bagi arsitektur CNN-LSTM dalam mengenali cyberbullying.

Tabel 9 menyajikan contoh representatif dari komentar yang mengalami kesalahan klasifikasi, dikategorikan menjadi False Positive (komentar non-bullying diprediksi sebagai bullying) dan False Negative (komentar bullying gagal terdeteksi). Berdasarkan analisis di Tabel 9, beberapa keterbatasan spesifik dari arsitektur CNN-LSTM dalam menangani data media sosial teridentifikasi. Pertama, Dominasi Fitur Lokal (Efek CNN) di mana kesalahan False Positive pada contoh no. 1 menunjukkan bahwa lapisan CNN sangat sensitif terhadap kata kunci negatif (seperti "anjing"). Meskipun lapisan LSTM bertugas menangkap konteks, fitur n-gram lokal yang diekstraksi oleh CNN terkadang terlalu dominan, sehingga satu kata kasar dapat menyebabkan kalimat netral diprediksi sebagai bullying. Kedua, Keterbatasan Pemahaman Bahasa Code-Mixing di mana model menunjukkan kesulitan dalam menangani kalimat yang sepenuhnya dalam Bahasa Inggris atau campuran (contoh no. 2). Hal ini mengindikasikan bahwa meskipun menggunakan FastText embeddings, representasi semantik untuk frasa asing yang panjang masih belum optimal dalam arsitektur ini. Ketiga, Kompleksitas Pragmatis dan Slang di mana kesalahan False Negative pada contoh no. 4 menyoroti tantangan dalam mendeteksi bullying yang disamarkan dalam slang atau kalimat sarkastik. Arsitektur CNN-LSTM mungkin kehilangan nuansa implisit ini karena operasi pooling di CNN mengurangi dimensi fitur, berpotensi menghilangkan informasi urutan kata yang halus namun penting untuk mendeteksi sarkasme.

Temuan kualitatif ini mengonfirmasi bahwa meskipun model CNN-LSTM memiliki kinerja statistik yang tinggi (F1 lebih besar dari 88%), pengembangan lebih lanjut perlu berfokus pada mekanisme penanganan ambiguitas kata dan pengayaan data pelatihan dengan variasi slang dan bahasa campuran yang lebih beragam. Analisis kualitatif ini memberikan wawasan berharga bahwa peningkatan kinerja di masa depan tidak hanya bergantung pada kompleksitas arsitektur, tetapi juga pada pengayaan data pelatihan untuk mencakup variasi linguistik yang lebih luas.

VII. KESIMPULAN

Penelitian ini mengusulkan dan mengevaluasi arsitektur hibrida CNN-LSTM untuk deteksi perundungan siber di media sosial X berbahasa Indonesia menggunakan dataset sebanyak 13.677 komentar. Penelitian ini memperoleh beberapa temuan kunci melalui eksperimen sistematis di tiga skenario. Pertama, mekanisme Early Stopping bersifat kritis untuk mencegah overfitting. Model yang dilatih tanpa mekanisme ini mengalami degradasi kinerja yang parah dengan akurasi menurun hingga 55,92%, sementara model dengan Early Stopping mencapai akurasi 88,38% dan AUC 0,9346. Kedua, embedding FastText yang telah dilatih sebelumnya meningkatkan kemampuan diskriminatif model, meningkatkan AUC menjadi 0,9559. Ketiga, studi ablasi komparatif mengungkapkan bahwa BiLSTM dengan Attention mencapai kinerja tertinggi di antara model pembelajaran mendalam dengan F1-Score 88,77%, meskipun hibrida CNN-LSTM dan CNN-BiGRU menawarkan efisiensi komputasi yang lebih baik dengan akurasi kompetitif masing-masing 88,38% dan 88,34%.

Benchmarking terhadap metode mutakhir menunjukkan bahwa model yang diusulkan secara signifikan mengungguli SVM klasik dengan margin lebih dari 10 poin persentase. IndoBERTweet mencapai akurasi tertinggi sebesar 91,59%, namun arsitektur CNN-LSTM yang diusulkan mencapai 97% dari kinerja tersebut dengan 95% lebih sedikit parameter (5,32 juta berbanding 110 juta), sehingga menjadikannya sangat cocok untuk skenario deployment dengan sumber daya terbatas. Penelitian ini memiliki keterbatasan yang mencakup pembatasan dataset pada media sosial X, pendekatan klasifikasi biner, dan tantangan dalam menangani pencampuran kode serta perundungan implisit

seperti sarkasme. Penelitian selanjutnya perlu memperluas dataset ke berbagai platform, mengimplementasikan klasifikasi multi-kelas untuk tingkat keparahan, mengeksplorasi pendekatan multimodal yang menggabungkan teks dengan analisis visual.

UCAPAN TERIMA KASIH

Penelitian ini dapat terlaksana berkat dukungan dan fasilitas yang disediakan oleh Universitas Muhammadiyah Sidoarjo. Penulis menyampaikan rasa terima kasih yang sebesar-besarnya kepada dosen pembimbing atas bimbingan yang sangat berharga, masukan yang konstruktif, serta wawasan ilmiah yang diberikan selama proses penelitian. Bimbingan tersebut sangat berperan penting dalam membentuk metodologi dan analisis studi ini.

Kami juga mengapresiasi tim peneliti dan rekan-rekan yang telah berkontribusi dalam proses pengumpulan dan anotasi data. Sumber daya komputasi dalam penelitian ini didukung oleh Google Colaboratory dengan akselerasi GPU Tesla T4, menggunakan kerangka kerja (framework) TensorFlow dan Keras. Karya ini didedikasikan untuk mendukung pengembangan lingkungan digital yang lebih aman di Indonesia.

REFERENSI

[1] Shubham Singh, “Twitter Statistics 2025: How Many People Use X [New Data].” Accessed: Aug. 20, 2025. [Online]. Available: https://www.demandsage.com/twitter-statistics/

[2] A. M. El Koshiry, E. H. I. Eliwa, T. A. El-Hafeez, and M. Khairy, “Detecting cyberbullying using deep learning techniques: a pre-trained glove and focal loss technique,” PeerJ Comput. Sci., vol. 10, pp. 1–33, 2024, doi: 10.7717/peerj-cs.1961.

[3] M. T. Hasan, M. A. E. Hossain, M. S. H. Mukta, A. Akter, M. Ahmed, and S. Islam, “A Review on Deep-Learning-Based Cyberbullying Detection,” Future Internet, vol. 15, no. 5, pp. 1–47, 2023, doi: 10.3390/fi15050179.

[4] L. Iryani et al., “Model of Cyberbullying Detection on Social Media Using Multi-Label Deep Learning: a Comparative Study,” JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer), vol. 10, no. 4, pp. 742–748, 2025, doi: 10.33480/jitk.v10i4.6004.

[5] M. D. Maulana and C. S. K. Aditya, “Perbandingan IndoBERT dan Bi-LSTM Dalam Mendeteksi Pelanggaran Undang-Undang ITE,” Science And Information Technology (SINTECH), vol. 8, no. 1, pp. 52–59, 2025, [Online]. Available: https://doi.org/10.31598

[6] D. D. N. Cahyo, R. Handayani, V. B. Lestari, and S. Febriani, “JITE (Journal of Informatics and Telecommunication Engineering) Sentiment Analysis of Public Opinion on Online Gambling Through,” vol. 9, no. July, pp. 99–115, 2025.

[7] V. Ramadhan, A. Pambudi, T. Informatika, and U. M. Sukabumi, “IMPLEMENTASI ALGORITMA CONVOLUTIONAL NEURAL NETWORK UNTUK MENGIDENTIFIKASI BERITA HOAKS BERBAHASA INDONESIA,” vol. 8, no. 5, pp. 10945–10952, 2024.

[8] I. A. Asqolani and E. B. Setiawan, “A Hybrid Deep Learning Approach Leveraging Word2Vec Feature Expansion for Cyberbullying Detection in Indonesian Twitter,” Ingenierie des Systemes d’Information, vol. 28, no. 4, pp. 887–895, 2023, doi: 10.18280/isi.280410.

[9] M. A. S. Nasution and E. B. Setiawan, “Enhancing Cyberbullying Detection on Indonesian Twitter: Leveraging FastText for Feature Expansion and Hybrid Approach Applying CNN and BiLSTM,” Revue d’Intelligence Artificielle, vol. 37, no. 4, pp. 929–936, 2023, doi: 10.18280/ria.370413.

[10] 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, 2025, doi: 10.52436/1.jutif.2025.6.2.4321.

[11] J. Faritha Banu et al., “Modeling of Hyperparameter Tuned Hybrid CNN and LSTM for Prediction Model,” Intelligent Automation and Soft Computing, vol. 33, no. 3, pp. 1393–1405, 2022, doi: 10.32604/iasc.2022.024176.

[12] S. Mitra, T. Tasnim, M. A. R. Islam, N. I. Khan, and M. S. Majib, “A Framework to Detect and Prevent Cyberbullying from Social Media by Exploring Machine Learning Algorithms,” 6th International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering, IC4ME2 2021, no. February 2023, 2021, doi: 10.1109/IC4ME253898.2021.9768450.

[13] R. Masbadi Hatullah Nurnaryo, M. Mulaab, I. Oktavia Suzanti, D. Abdul Fatah, A. D. Cahyani, and F. Ayu Mufarroha, “Deteksi Cyberbullying Pada Data Tweet Menggunakan Metode Random Forest Dan Seleksi Fitur Information Gain,” Jurnal Simantec, vol. 11, no. 1, pp. 33–40, 2022, doi: 10.21107/simantec.v11i1.17256.

[14] H. Santoso, R. A. Putri, and S. Sahbandi, “Deteksi Komentar Cyberbullying pada Media Sosial Instagram Menggunakan Algoritma Random Forest,” Jurnal Manajemen Informatika (JAMIKA), vol. 13, no. 1, pp. 62–72, 2023, doi: 10.34010/jamika.v13i1.9303.

[15] A. J. Andika, Y. Kristian, and E. I. Setiawan, “Deteksi 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.

[16] D. Harish, M. Manimaran, V. P. Jayashakthi, and M. Alamelu, “Automatic Detection of Cyberbullying on Social Media Using Machine Learning,” 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation, ICAECA 2023, vol. 13, no. 3, pp. 185–189, 2023, doi: 10.1109/ICAECA56562.2023.10201149.

[17] A. F. Alqahtani and M. Ilyas, “An Ensemble-Based Multi-Classification Machine Learning Classifiers Approach to Detect Multiple Classes of Cyberbullying,” Mach. Learn. Knowl. Extr., vol. 6, no. 1, pp. 156–170, 2024, doi: 10.3390/make6010009.

[18] S. R. Ahmad, N. Insani, and M. Salim, “Analysis of Cyberbullying on Social Media Using A Comparison of Naïve Bayes, Random Forest, and SVM Algorithms,” Jurnal Teknologi Informasi dan Pendidikan, vol. 17, no. 1, pp. 75–86, 2024, doi: 10.24036/jtip.v17i1.807.

[19] F. Akar, “Performance Analysis of NLP-Based Machine Learning Algorithms in Cyberbullying Detection,” Erzincan Üniversitesi Fen Bilimleri Enstitüsü Dergisi, vol. 17, no. 2, pp. 445–459, 2024, doi: 10.18185/erzifbed.1474112.

[20] V. Tapaskar and M. C. T. J, “INTERNATIONAL JOURNAL OF A Multi-Algorithm Approach for Cyberbullying Detection,” vol. 7, no. 10, 2024, doi: 10.15680/IJMRSET.2024.0710027.

[21] F. Akter, F. Rabbi, and R. R. Chowdhury, “Cyberbullying Detection on Social Media Platforms Utilizing Different Machine Learning Approaches,” vol. 186, no. 61, pp. 40–50, 2025.

[22] B. S. Purkayastha, M. Rahman, and T. I. Talukdar, “Advancing Cyberbullying Detection : A Hybrid Machine Learning and Deep Learning Framework for Social Media Analysis,” vol. 2, no. Iceis, pp. 348–355, 2025, doi: 10.5220/0013436200003929.

[23] T. T. Prama, J. F. Amrin, Md. M. Anwar, and I. H. Sarker, “AI Enabled User-Specific Cyberbullying Severity Detection with Explainability,” pp. 1–25, 2025, [Online]. Available: http://arxiv.org/abs/2503.10650

[24] M. A. Rosid, D. Siahaan, and A. Saikhu, “Sarcasm Detection in Indonesian-English Code-Mixed Text Using Multihead Attention-Based Convolutional and Bi-Directional GRU,” IEEE Access, no. June, pp. 137063–137079, 2024, doi: 10.1109/ACCESS.2024.3436107.
[25] C. D. I. Smote, “Analisis komparatif klasifikasi sentimen pengguna aplikasi investasi menggunakan algoritma hybrid cnn-lstm, cnn-gru dengan implementasi smote,” vol. 6, pp. 112–118, 2025..