



Development of an Explainable Multi-Task Transformer Framework for Detecting and Classifying the Severity of Cyberbullying in Social Media Comments

Pengembangan Framework Explainable Multi-Task Transformer untuk Deteksi dan Klasifikasi Tingkat Keparahan Cyberbullying pada Komentar Media Sosial

Elma Salsabella^{1*}, Sampurna Dadi Riskiono²

^{1,2}Program Studi Sistem Informasi, Fakultas Teknik dan Ilmu Komputer,
Universitas Teknokrat Indonesia, Indonesia

E-Mail: ¹elma_salsabella@teknokrat.ac.id, ²sampurna.go@teknokrat.ac.id

Received Apr 16th 2026; Revised May 23th 2026; Accepted Jun 09th 2026; Available Online Jun 21th 2026

Corresponding Author: Elma Salsabella

Copyright © 2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

Cyberbullying on social media has become a serious threat to the mental health of internet users in Indonesia. However, research on automatic Indonesian cyberbullying detection remains limited and generally focuses only on binary classification without considering severity levels and model interpretability. This study proposes an IndoBERT-based Multi-Task Learning (MTL) approach to simultaneously perform cyberbullying detection and four-class severity classification within a unified model. The dataset consists of 4,795 Indonesian social media comments obtained through a rule-based hybrid auto-labeling approach combined with semi-supervised refinement. Manual validation and consistency checking were conducted to improve label quality before model training. Experimental results show that the proposed model achieved a weighted F1-score of 0.9374 for cyberbullying detection and 0.9027 for severity classification, consistently outperforming the Support Vector Machine (SVM) and Bidirectional Long Short-Term Memory (BiLSTM) baselines. The ablation study demonstrated that the multi-task approach and preprocessing significantly contributed to performance improvement. In addition, integrating Explainable Artificial Intelligence (XAI) with Local Interpretable Model-agnostic Explanations (LIME) revealed that toxic words were dominant features influencing model predictions. The findings indicate that the proposed approach is effective and robust for detecting cyberbullying in noisy social media data.

Keywords: Cyberbullying, Multi-Task Learning, IndoBERT, Explainable AI, LIME

Abstrak

Cyberbullying di media sosial menjadi ancaman serius bagi kesehatan mental pengguna internet di Indonesia. Namun, penelitian deteksi otomatis cyberbullying berbahasa Indonesia masih terbatas dan umumnya hanya berfokus pada klasifikasi biner tanpa mempertimbangkan tingkat keparahan (*severity*) serta interpretabilitas model. Penelitian ini mengusulkan pendekatan *Multi-Task Learning* (MTL) berbasis IndoBERT untuk melakukan deteksi cyberbullying dan klasifikasi *severity* dengan 4 kelas secara simultan dalam satu model. Dataset terdiri dari 4.795 komentar media sosial berbahasa Indonesia yang diperoleh melalui pendekatan *auto-labeling hybrid* berbasis aturan dan penyempurnaan *semi-supervised*. Validasi manual dan pengecekan konsistensi dilakukan untuk meningkatkan kualitas label sebelum pelatihan model. Hasil eksperimen menunjukkan bahwa model memperoleh *weighted F1-score* sebesar 0,9374 untuk deteksi cyberbullying dan 0,9027 untuk klasifikasi *severity*, serta mengungguli *baseline* Support Vector Machine (SVM) dan Bidirectional Long Short-Term Memory (BiLSTM). *Ablation study* menunjukkan bahwa pendekatan *multi-task* dan *preprocessing* memberikan kontribusi signifikan terhadap peningkatan performa. Selain itu, integrasi *Explainable Artificial Intelligence* (XAI) menggunakan Local Interpretable Model-agnostic Explanations (LIME) menunjukkan bahwa kata-kata toksik menjadi fitur dominan dalam prediksi model. Penelitian ini membuktikan bahwa pendekatan yang diusulkan efektif dan *robust* untuk mendeteksi cyberbullying pada data media sosial yang bersifat *noisy*.

Kata Kunci: Cyberbullying, Multi-Task Learning, IndoBERT, Explainable AI, LIME

1. PENDAHULUAN

Perkembangan pesat media sosial di Indonesia membawa dampak positif dalam konektivitas dan pertukaran informasi. Namun, fenomena *cyberbullying* menjadi permasalahan yang semakin mengkhawatirkan. Laporan UNICEF tahun 2023 menunjukkan bahwa Indonesia menempati peringkat tertinggi di Asia Tenggara untuk kasus *cyberbullying* pada remaja [1]. Platform seperti YouTube, Instagram, Twitter/X, dan TikTok menjadi medium utama perundungan daring [2]. *Cyberbullying* didefinisikan sebagai tindakan agresif yang dilakukan secara sengaja melalui media elektronik terhadap individu yang sulit membela diri [3]. Berbeda dengan bullying tradisional, *cyberbullying* memiliki karakteristik unik berupa anonimitas pelaku, jangkauan audiens yang luas, serta persistensi konten digital yang sulit dihapus [4].

Deteksi otomatis *cyberbullying* menjadi kebutuhan mendesak mengingat volume konten media sosial yang sangat besar, sehingga moderasi manual tidak memungkinkan untuk diterapkan secara menyeluruh. Pendekatan berbasis Natural Language Processing (NLP) menawarkan solusi yang skalabel [5]. Metode awal menggunakan *machine learning* tradisional seperti Naive Bayes, Support Vector Machine (SVM), dan Random Forest dengan representasi *Bag-of-Words* atau *Term Frequency-Inverse Document Frequency* (TF-IDF) [6], namun memiliki keterbatasan dalam menangkap konteks semantik. Kemajuan *deep learning* melalui Convolutional Neural Network (CNN) dan Long Short-Term Memory (LSTM) meningkatkan performa klasifikasi teks bullying [7], sementara arsitektur berbasis *attention mechanism* turut memperkuat kemampuan dalam menangkap fitur kontekstual [8]. Meskipun demikian, model-model tersebut masih bergantung pada representasi kata statis sehingga kurang optimal dalam memahami makna kontekstual yang kompleks. Sebagai respons terhadap keterbatasan tersebut, model bahasa pra-latih (*pre-trained language models*) seperti Bidirectional Encoder Representations from Transformers (BERT) [9], membawa paradigma baru dalam NLP dengan kemampuan memahami konteks secara *bidirectional*. Untuk bahasa Indonesia, IndoBERT merupakan model bahasa pra-latih berbasis arsitektur BERT yang dilatih pada korpus bahasa Indonesia berskala besar dan telah terbukti efektif untuk berbagai tugas NLP, termasuk deteksi *cyberbullying* [10]. Berbagai studi komparatif menunjukkan keunggulan IndoBERT dibandingkan model umum dalam memahami karakteristik linguistik bahasa Indonesia [11].

Lebih lanjut, *Multi-Task Learning* (MTL) merupakan paradigma yang melatih model pada beberapa tugas terkait secara simultan sehingga memungkinkan pembelajaran representasi yang lebih general melalui *shared parameters* [12]. Dalam konteks deteksi *cyberbullying*, MTL berpotensi menangkap hubungan semantik antara keberadaan bullying dan tingkat keparahannya (*severity*). Di sisi lain, *Explainable Artificial Intelligence* (XAI) menjadi aspek penting dalam moderasi konten otomatis untuk membangun kepercayaan pengguna serta memenuhi tuntutan regulasi [13]. Salah satu metode yang banyak digunakan adalah Local Interpretable Model-agnostic Explanations (LIME), yang mampu memberikan penjelasan lokal terhadap prediksi model melalui pendekatan *perturbation* pada input [14]. Meskipun penelitian deteksi *cyberbullying* telah berkembang pesat, sebagian besar studi masih berfokus pada teks berbahasa Inggris [15]. Penelitian untuk bahasa Indonesia masih relatif terbatas dan menghadapi tantangan unik seperti penggunaan bahasa gaul (*slang*), *code-mixing*, serta ekspresi budaya lokal [16]. Selain itu, sebagian besar penelitian yang ada hanya melakukan klasifikasi biner (*bullying* vs *non-bullying*) tanpa mempertimbangkan tingkat keparahan, serta belum mengintegrasikan aspek interpretabilitas model secara memadai.

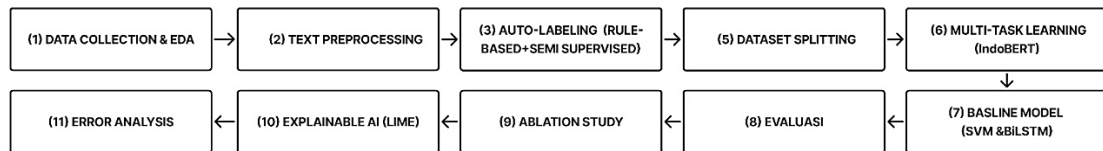
Secara komparatif, penelitian terdahulu menunjukkan variasi pendekatan yang digunakan. Obaid dkk. (2023) [17] menggunakan CNN, LSTM, dan BERT pada data berbahasa Inggris dengan fokus pada *severity* tanpa *multi-task learning* maupun XAI. Jaradat dkk. (2025) [18] mengkaji pendekatan *deep learning* pada data multibahasa tanpa mempertimbangkan *severity*, MTL, maupun interpretabilitas. Pamungkas dkk. (2023) [19] dan Nurrahmi & Nurjanah (2021) [20] berfokus pada bahasa Indonesia, namun masih menggunakan pendekatan *machine learning* atau IndoBERT secara *single-task* tanpa mempertimbangkan *severity* dan XAI. Rathnayake dkk. (2024) [21] telah mengimplementasikan BERT dengan MTL, tetapi belum mengakomodasi klasifikasi *severity* maupun interpretabilitas model. Sementara itu, Awal dkk. (2023) [22] mengkaji *severity* menggunakan BERT tanpa MTL dan XAI. Meskipun berbagai pendekatan telah diusulkan, terdapat tiga keterbatasan utama dalam literatur saat ini. Pertama, belum adanya pendekatan terpadu yang menggabungkan klasifikasi *cyberbullying* dan *severity* dalam kerangka MTL untuk bahasa Indonesia. Kedua, sebagian besar penelitian menggunakan dataset berlabel manual yang relatif bersih, sehingga belum mencerminkan kondisi nyata data media sosial yang cenderung *noisy*. Ketiga, masih minimnya integrasi XAI untuk mengevaluasi interpretabilitas model dalam konteks bahasa informal.

Berdasarkan kesenjangan tersebut, penelitian ini bertujuan untuk: (1) mengembangkan sistem deteksi *cyberbullying* berbasis MTL menggunakan IndoBERT yang mampu mengklasifikasikan teks ke dalam kategori *bullying/non-bullying* dan tingkat keparahan (*severity*); (2) mengevaluasi performa model melalui perbandingan dengan *baseline* serta *ablation study*; (3) mengimplementasikan XAI menggunakan LIME untuk meningkatkan transparansi keputusan model; dan (4) melakukan analisis kesalahan untuk mengidentifikasi keterbatasan model. Berbeda dengan penelitian sebelumnya, kontribusi utama studi ini tidak hanya terletak pada kombinasi metode, tetapi pada integrasi terpadu antara MTL, klasifikasi *severity*, dan XAI yang dievaluasi pada dataset bahasa Indonesia dengan karakteristik *noisy* dan semi-otomatis (*auto-*

labeled). Hingga saat ini, belum terdapat studi yang secara sistematis mengevaluasi efektivitas pendekatan MTL dalam skenario *noisy auto-labeled dataset* untuk deteksi *cyberbullying* bahasa Indonesia, khususnya dalam konteks klasifikasi hierarkis antara keberadaan bullying dan tingkat keparahannya. Selain itu, penelitian ini juga memberikan kontribusi metodologis melalui pendekatan *hybrid auto-labeling* yang menggabungkan *rule-based labeling* dan *self-training*, yang masih jarang dieksplorasi dalam domain ini. Integrasi XAI memungkinkan evaluasi tidak hanya pada performa model, tetapi juga pada konsistensi dan relevansi fitur linguistik yang dipelajari model, sehingga memberikan wawasan baru terkait interpretabilitas model pada teks dengan karakteristik *slang* dan *code-mixing*. Dengan demikian, penelitian ini berkontribusi tidak hanya pada peningkatan performa, tetapi juga pada *robustness* terhadap *noise*, generalisasi model, serta transparansi dalam sistem moderasi konten otomatis.

2. METODE PENELITIAN

Sistem yang diusulkan terdiri dari delapan tahap utama, yaitu pengumpulan data dan *Exploratory Data Analysis* (EDA), *preprocessing* teks, *auto-labeling* berbasis *rule-based*, pembagian dataset menggunakan *stratified sampling*, *semi-supervised refinement* pada data latih untuk menghindari *data leakage*, pemodelan menggunakan pendekatan *Multi-Task Learning* (MTL) berbasis IndoBERT dengan dua model *baseline*, yaitu Support Vector Machine (SVM) dan Bidirectional Long Short-Term Memory (BiLSTM) sebagai pembanding, evaluasi dan perbandingan model termasuk *ablation study*, serta penerapan *Explainable Artificial Intelligence* (XAI) menggunakan Local Interpretable Model-agnostic Explanations (LIME) untuk meningkatkan interpretabilitas model yang dilengkapi dengan analisis kesalahan [23]. Gambar 1 mengilustrasikan arsitektur keseluruhan sistem.



Gambar 1. mengilustrasikan arsitektur keseluruhan sistem.

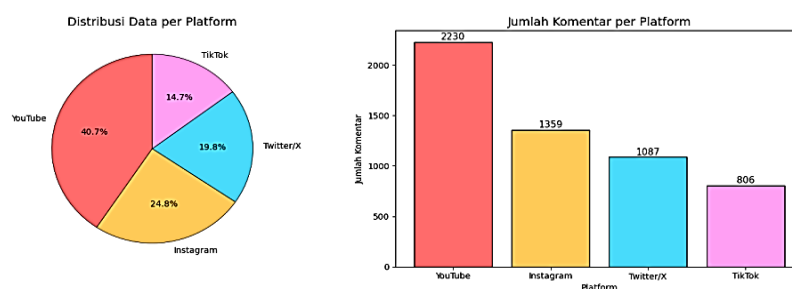
2.1. Pengumpulan Data & EDA

Dataset penelitian ini berasal dari komentar media sosial berbahasa Indonesia yang dikumpulkan dari YouTube, Instagram, Twitter/X, dan TikTok menggunakan layanan Apify melalui API. Pengumpulan data dilakukan berdasarkan kata kunci terkait *cyberbullying*, seperti “bodoh”, “tolol”, “goblok”, “anjir”, “sampah”, “ga jelas”, serta frasa kontekstual seperti “konten cringe” dan “komen pedas”. Pendekatan ini memungkinkan diperolehnya data yang beragam, meskipun berpotensi menimbulkan bias terhadap komentar negatif. Data disimpan dalam format JSON/CSV dengan atribut *Date*, *UserName*, *Comment*, dan *Platform*. Dari total 5.482 komentar yang dikumpulkan, sebanyak 543 data duplikat dan kosong dihapus pada tahap *data cleaning*, sehingga diperoleh 4.795 komentar valid.[24] Contoh data ditampilkan pada Tabel 1.

Tabel 1. Contoh Dataset Komentar

No	Date	UserName	Comment	Platform
1	2024-10-02	rizky_pratama	Dsr bodoh, gak punya otak!	Twitter/X
2	2024-10-03	dinaa_ayu	Konten sampah kayak gini masih ada aja!!	TikTok
3	2024-10-04	andi_setiawan	Ngomong aja ga jelas, mending diam	YouTube
4.795	2024-10-05	putri_lestari	Kontennya cukup menarik, tapi bisa improve	Instagram

Berdasarkan Tabel 1, dataset memiliki struktur yang terdiri dari tanggal, username, isi komentar, dan platform asal, serta mencerminkan variasi komentar bullying dan non-bullying. Distribusi jumlah data pada setiap platform dapat dilihat pada Gambar 2.



Gambar 2. Distribusi Per Platform

Distribusi data pada setiap platform ditunjukkan pada Gambar 2, dengan dominasi YouTube (2.230), diikuti Instagram (1.359), Twitter/X (1.087), dan TikTok (806). Ketidakseimbangan ini mengindikasikan adanya bias platform, di mana model berpotensi dipengaruhi oleh karakteristik bahasa pada platform dominan. Selain itu, penggunaan *keyword-based scraping* berpotensi menimbulkan *selection bias*, karena komentar yang bersifat implisit atau sarkastik mungkin tidak terambil. Secara keseluruhan, dataset memiliki karakteristik *noisy* dan heterogen, sehingga memerlukan pendekatan pemodelan yang robust, seperti Multi-Task Learning, untuk meningkatkan kemampuan generalisasi pada data dunia nyata.

2.2. Preprocessing Teks

Preprocessing merupakan tahap penting mengingat karakteristik teks media sosial berbahasa Indonesia yang cenderung informal dan mengandung banyak *noise* [25]. Tahapan yang dilakukan meliputi *lowercase conversion*, penghapusan URL, emoji Unicode, angka, dan simbol, reduksi huruf berulang maksimal dua kali, normalisasi *slang* menggunakan kamus berisi lebih dari 110 entri yang dikurasi secara manual, serta *stopword removal* menggunakan *library* Sastrawi. Kamus normalisasi mencakup berbagai kategori, seperti negasi, kata ganti, *intensifier*, *slang* media sosial, dan kata kasar. Contoh hasil *preprocessing* ditampilkan pada Tabel 2, yang menunjukkan bahwa teks berhasil dibersihkan dari *noise* seperti singkatan, tanda baca berlebih, dan kata tidak baku, sehingga menghasilkan representasi teks yang lebih terstruktur.

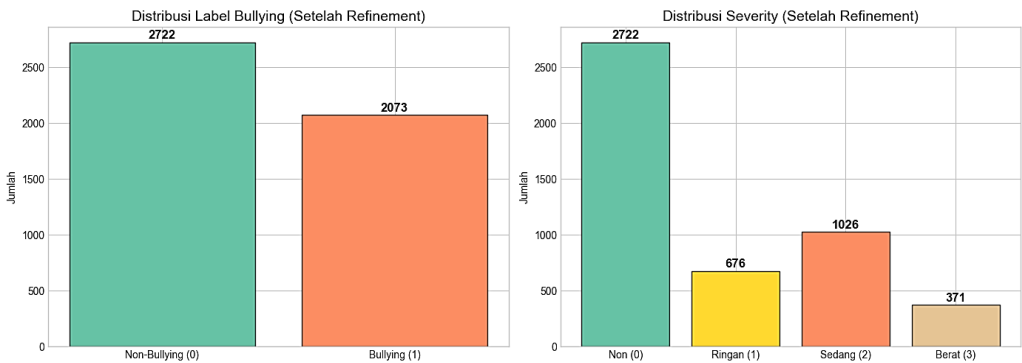
Tabel 2. Contoh Hasil *Preprocessing*

Comment (Asli)	Hasil <i>Preprocessing</i>
Dsr BODOH, gak punya otak!	dasar bodoh tidak punya otak
Konten sampah kayak gini masih ada aja!!	konten sampah seperti ini masih ada
Ngomong aja ga jelas, mending diam	ngomong saja tidak jelas mending diam
Kontennya cukup menarik, tapi bisa improve	konten cukup menarik tapi bisa improve

Penggunaan *stopword removal* bertujuan untuk mengurangi kata-kata umum yang kurang informatif dalam proses klasifikasi. Namun, dalam penelitian ini, *stopword* yang memiliki peran penting dalam konteks sentimen dan makna, seperti kata negasi (misalnya “tidak” dan “bukan”), tetap dipertahankan agar tidak menghilangkan informasi semantik yang krusial. Selain itu, *stemming* tidak digunakan karena berpotensi menghilangkan nuansa makna yang penting dalam konteks *cyberbullying*, terutama pada kata-kata yang mengekspresikan emosi atau intensitas. Proses *tokenizing* juga tidak dilakukan secara eksplisit karena IndoBERT menggunakan mekanisme *subword tokenization*, sehingga mampu menangani variasi morfologis dan *slang* dengan lebih efektif. Pada model *baseline* Support Vector Machine (SVM), representasi fitur teks menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF). Sementara itu, pada model IndoBERT tidak digunakan TF-IDF karena representasi teks diperoleh secara langsung melalui *contextual embedding* berbasis *transformer*.

2.3. Pelabelan Data

Keterbatasan dataset berlabel diatasi dengan pendekatan *auto-labeling* berbasis aturan. Kata kunci dikelompokkan ke dalam tiga tingkat *severity*, yaitu berat, sedang, dan ringan, dengan pelabelan yang dilakukan menggunakan *substring matching* berdasarkan prioritas tingkat keparahan tertinggi. Untuk meningkatkan kualitas label, diterapkan pendekatan *semi-supervised self-training* menggunakan SVM (LinearSVC) berbasis TF-IDF. Proses ini dilakukan hanya pada data latih untuk menghindari *data leakage* [26], dengan dua iterasi dan *confidence threshold* sebesar 0,7. Distribusi label ditunjukkan pada Gambar 3.



Gambar 3. Distribusi Label

Distribusi data menunjukkan bahwa kelas *non-bullying* berjumlah 2.722 dan *bullying* 2.073 komentar. Perbandingan ini masih tergolong relatif seimbang karena selisih antar kelas tidak terlalu jauh (sekitar 57% :

43%), sehingga model umumnya masih dapat belajar dengan baik tanpa teknik penyeimbangan tambahan. Oleh karena itu, penerapan SMOTE tidak wajib dilakukan pada klasifikasi biner bullying. Namun, pada distribusi *severity* terdapat ketimpangan yang lebih jelas, terutama pada kelas *berat* yang hanya berjumlah 371 data dibandingkan dengan kelas *non* sebanyak 2.722 data. Ketidakseimbangan ini berpotensi menyebabkan model lebih bias terhadap kelas mayoritas dan menurunkan kemampuan deteksi pada kelas minoritas. Dalam kondisi tersebut, penggunaan SMOTE dapat dipertimbangkan untuk meningkatkan representasi kelas minoritas dalam klasifikasi tingkat *severity*.

2.4. Data Splitting

Data dibagi menggunakan *stratified split* dengan proporsi 70:10:20 untuk data latih, validasi, dan uji. Teknik ini digunakan untuk menjaga keseimbangan distribusi kelas *bullying* dan *non-bullying* pada setiap subset. Data latih digunakan untuk pelatihan model, data validasi untuk tuning dan pemantauan overfitting, sedangkan data uji digunakan untuk evaluasi akhir[27]. Rincian pembagian dataset disajikan pada Tabel 3.

Tabel 3. Pembagian Dataset

Subset	Total	Non-Bullying	Bullying	Rasio
Training	3.356	1.909	1.447	0,758
Validasi	480	271	209	0,771
Testing	959	542	417	0,769

Tabel 3 menunjukkan distribusi data yang tetap terjaga antarkelas pada setiap subset. Verifikasi *data leakage* juga dilakukan untuk memastikan tidak terdapat duplikasi teks antarsubset.

2.5. Arsitektur Model Multi-Task Learning

Model yang diusulkan menggunakan pendekatan MTL dengan indobenchmark/indobert-base-p1 (IndoBERT) sebagai *shared encoder*. Encoder berbasis Transformer dengan *hidden size* 768 menghasilkan representasi kontekstual, di mana token [CLS] digunakan sebagai representasi global untuk dua tugas, yaitu deteksi *cyberbullying* (klasifikasi biner) dan klasifikasi tingkat keparahan (multi-kelas). Pendekatan ini dipilih karena kedua tugas saling berkaitan, sehingga memungkinkan pembelajaran fitur bersama yang lebih efisien dan meningkatkan generalisasi model. Untuk tugas deteksi *cyberbullying*, arsitektur *classifier head* terdiri dari Dropout(0.3), Linear(768,256), ReLU, Dropout(0.3), dan Linear(256,1). Output berupa logit yang dikonversi menjadi probabilitas menggunakan fungsi sigmoid, dengan keputusan klasifikasi berdasarkan *threshold* 0,5[28]. Fungsi loss yang digunakan adalah *Binary Cross-Entropy with Logits Loss* yang dirumuskan pada persamaan (1).

$$L_{\text{bullying}} = -\frac{1}{N} \sum_{i=1}^N (y_i \log \sigma(\hat{y}_i) + (1 - y_i) \log(1 - \sigma(\hat{y}_i))) \quad (1)$$

Penggunaan loss ini memberikan stabilitas numerik karena menggabungkan fungsi sigmoid dan cross-entropy dalam satu operasi. Untuk tugas klasifikasi tingkat keparahan, arsitektur *classifier head* terdiri dari Dropout(0.3), Linear(768,256), ReLU, Dropout(0.3), dan Linear(256,4), dengan output berupa vektor logit yang dinormalisasi menggunakan softmax. Fungsi loss yang digunakan adalah *categorical cross-entropy* dirumuskan pada persamaan (2).

$$L_{\text{severity}} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{i,c} \log \left(\frac{e^{\hat{y}_{i,c}}}{\sum_{j=1}^C e^{\hat{y}_{i,j}}} \right) \quad (2)$$

dengan $C=4$ yang merepresentasikan kelas non, ringan, sedang, dan berat. Fungsi softmax memastikan distribusi probabilitas antar kelas. Karena kedua tugas dilatih secara simultan, digunakan fungsi loss gabungan untuk mengontrol kontribusi masing-masing tugas dirumuskan pada persamaan (3).

$$L_{\text{total}} = \alpha L_{\text{bullying}} + \beta L_{\text{severity}} \quad (3)$$

dengan $\alpha=0,5$ dan $\beta=0,5$ untuk memberikan kontribusi yang seimbang. Proses pelatihan menggunakan konfigurasi hyperparameter sebagai berikut pada Tabel 4. Strategi *early stopping* digunakan untuk mencegah *overfitting*, sedangkan *gradient clipping* menjaga stabilitas pelatihan. Secara keseluruhan, model memiliki sekitar 124 juta parameter trainable yang mencerminkan kompleksitas arsitektur berbasis Transformer.

2.6. Model Baseline

Penelitian ini menggunakan dua model baseline, yaitu Support Vector Machine (SVM) dan Bidirectional Long Short-Term Memory (BiLSTM), sebagai pembandingan kinerja model yang diusulkan.

SVM menggunakan kernel *Radial Basis Function* (RBF) dengan parameter $C = 1,0$ dan $\gamma = scale$. Fitur teks direpresentasikan menggunakan *TF-IDF Vectorizer* dengan $max_features = 10.000$ dan $n_gram (1,2)$ [29]. Pembobotan *TF-IDF* dihitung pada persamaan (4).

$$TF-IDF(t, d) = TF(t, d) \times IDF(t) \quad (4)$$

Model SVM dilatih secara terpisah untuk dua tugas, yaitu deteksi *cyberbullying* dan klasifikasi *severity*, agar masing-masing tugas dioptimalkan secara independen. Model kedua adalah *BiLSTM* dua lapisan dengan total 4.567.429 parameter *trainable*. Model ini menggunakan *embedding* berdimensi 128, *hidden dimension* 128, dan *dropout* 0,3 untuk mengurangi *overfitting*. Mekanisme LSTM terdiri dari *forget gate*, *input gate*, dan *output gate* yang mengatur aliran informasi pada setiap *timestep*. Arsitektur *BiLSTM* memproses teks dari dua arah (*forward* dan *backward*) untuk menangkap konteks yang lebih kaya. Proses pelatihan menggunakan *optimizer Adam* dengan *learning rate* 10^{-3} , serta dua *classification head* terpisah untuk tugas *cyberbullying* dan *severity*.

Tabel 4. Konfigurasi Hyperparameter

Parameter	Nilai
Optimizer	AdamW
Learning rate	2×10^{-5}
Weight decay	0,01
Batch size	16
Max sequence length	128 token
Maksimum epoch	10
Early stopping patience	3
Gradient clipping	Max norm = 1,0
LR scheduler	ReduceLROnPlateau (patience = 2, factor = 0,5)

2.7. Explainable AI

Untuk meningkatkan interpretabilitas model, digunakan metode Local Interpretable Model-agnostic Explanations (LIME) [16] yang mampu menjelaskan prediksi secara independen dari arsitektur model. Dua *explainer* diinisialisasi secara terpisah untuk masing-masing tugas, yaitu deteksi *cyberbullying* dan klasifikasi tingkat *severity*, sehingga interpretasi yang dihasilkan lebih spesifik. Konfigurasi LIME meliputi $num_features = 10$ untuk menampilkan kata-kata yang paling berpengaruh, serta $num_samples = 500$ untuk membangun aproksimasi model lokal di sekitar data yang dianalisis. Selain itu, dilakukan agregasi global terhadap 100 sampel *test set* dengan menghitung rata-rata *feature importance* setiap kata, guna mengidentifikasi pola kontribusi fitur secara keseluruhan [30].

2.8. Evaluasi Model

Evaluasi kinerja model dilakukan menggunakan enam metrik klasifikasi standar [31]. *Accuracy* mengukur proporsi prediksi yang benar terhadap seluruh data, dirumuskan pada persamaan (5)-(10).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

di mana TP, TN, FP, dan FN masing-masing merepresentasikan True Positive, True Negative, False Positive, dan False Negative. Selanjutnya, *Precision* mengukur ketepatan prediksi positif, sedangkan *Recall* mengukur kemampuan model dalam menangkap seluruh data positif:

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

Keseimbangan antara *Precision* dan *Recall* direpresentasikan melalui *F1-Score* sebagai rata-rata harmonik keduanya:

$$F1-Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

Untuk kasus klasifikasi multi-kelas, digunakan dua variasi *F1-Score*. *F1-Score Weighted* mempertimbangkan proporsi jumlah sampel pada setiap kelas:

$$F1_{\text{weighted}} = \sum_{c=1}^C \frac{n_c}{N} \cdot F1_c \quad (9)$$

Sementara itu, *F1-Score Macro* dihitung sebagai rata-rata tanpa pembobotan antar kelas:

$$F1_{\text{macro}} = \frac{1}{C} \sum_{c=1}^C F1_c \quad (10)$$

Dalam penelitian ini, *F1-Score Macro* dipilih sebagai metrik utama karena lebih representatif untuk data yang tidak seimbang. Distribusi kelas *severity* yang tidak proporsional menuntut evaluasi yang memberikan bobot yang sama pada setiap kelas, sehingga performa model pada kelas minoritas tetap diperhatikan

3. HASIL DAN PEMBAHASAN

Hasil evaluasi model Multi-Task IndoBERT serta perbandingannya dengan *baseline* (SVM dan BiLSTM) pada tugas deteksi *cyberbullying* dan klasifikasi *severity* disajikan menggunakan metrik *Accuracy*, *Precision*, *Recall*, dan *F1-Score* untuk mengevaluasi kinerja model secara komprehensif. Analisis komparatif dan *ablation study* dilakukan untuk menilai kontribusi arsitektur *multi-task* dan *preprocessing*, sementara interpretasi model menggunakan LIME untuk mengidentifikasi fitur kata yang berpengaruh terhadap prediksi.

3.1. Hasil Training Multi-Task IndoBERT

Proses training dilakukan dengan strategi early stopping berdasarkan *validation loss*. Model mencapai best validation loss sebesar 0,2500 pada epoch ke-2 dan training dihentikan pada epoch ke-5 karena tidak terjadi perbaikan selama tiga epoch berturut-turut. Dinamika training disajikan pada Tabel 5.

Tabel 5. Dinamika Training Multi-Task IndoBERT

Epoch	Train Loss	Val Loss	Val Acc (Bullying)	Val Acc (Severity)
1	0,6412	0,4589	89,58%	78,12%
2	0,3352	0,2500	94,37%	89,38%
3	0,1853	0,2664	94,37%	90,00%
4	0,1073	0,2935	95,00%	91,04%
5	0,0744	0,3260	94,79%	92,29%

Berdasarkan Tabel 5, *train loss* menunjukkan penurunan yang konsisten, yang menandakan bahwa model mampu mempelajari pola pada data latih dengan baik. Namun, peningkatan *validation loss* setelah epoch ke-2 mengindikasikan bahwa overfitting mulai terjadi. Penerapan *early stopping* terbukti efektif untuk menghentikan training pada titik optimal sebelum performa mengalami degradasi lebih lanjut. Model terbaik (epoch ke-2) selanjutnya digunakan untuk mengevaluasi pada test set dan dibandingkan dengan model *baseline*.

3.2. Perbandingan Model

Perbandingan kinerja antara model yang diusulkan, yaitu *Multi-Task IndoBERT*, dengan model *baseline* (SVM dan BiLSTM) pada tugas deteksi *cyberbullying* dan klasifikasi tingkat *severity*. Evaluasi dilakukan pada *test set* menggunakan metrik *Accuracy*, *Precision*, *Recall*, serta *F1-Score* untuk memberikan gambaran performa model secara komprehensif. Perbandingan performa model disajikan pada Tabel 6.

Tabel 6. Perbandingan Performa Model pada Test Set

Model	Task	Accuracy	Precision (W)	Recall (W)	F1 (W)	F1 (Macro)
Multi-Task IndoBERT	Bullying	0,9374	0,9374	0,9374	0,9374	0,9363
Multi-Task IndoBERT	Severity	0,9030	0,9058	0,9030	0,9027	0,8689
SVM (TF-IDF)	Bullying	0,8457	0,8494	0,8457	0,8436	0,8394
SVM (TF-IDF)	Severity	0,7414	0,7752	0,7414	0,7031	0,5625
LSTM	Bullying	0,8519	0,8519	0,8519	0,8519	0,8494
LSTM	Severity	0,7675	0,7632	0,7675	0,7647	0,6577

Berdasarkan Tabel 6, *Multi-Task IndoBERT* secara konsisten menunjukkan performa terbaik pada seluruh metrik dan kedua tugas. Pada deteksi *cyberbullying*, model mencapai *F1-Score weighted* sebesar 0,9374, meningkat sekitar 9 poin dibandingkan dengan SVM dan 8 poin dibandingkan dengan LSTM. Sementara itu, pada klasifikasi *severity*, peningkatan performa lebih signifikan dengan selisih hingga sekitar 20 poin terhadap SVM dan 14 poin terhadap LSTM. Hal ini menunjukkan bahwa pendekatan *multi-task learning* mampu memanfaatkan keterkaitan antartugas untuk menghasilkan representasi yang lebih efektif.

Untuk memperoleh pemahaman yang lebih mendalam, analisis dilanjutkan pada tingkat per-kelas untuk tugas deteksi *cyberbullying*, sebagaimana disajikan pada Tabel 7.

Performa per-kelas pada Tabel 7 menunjukkan keseimbangan yang baik antara *precision* dan *recall* pada kedua kelas. Hal ini mengindikasikan bahwa model mampu meminimalkan kesalahan klasifikasi baik *false positive* maupun *false negative*, yang merupakan aspek penting dalam sistem moderasi konten otomatis. Analisis selanjutnya difokuskan pada performa per-kelas untuk klasifikasi tingkat *severity*, pada Tabel 8.

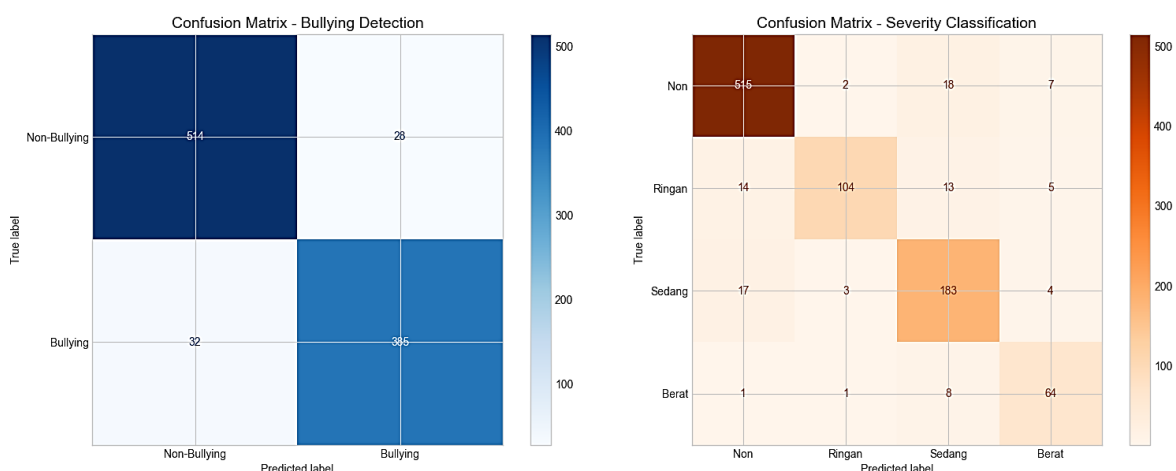
Tabel 7. Performa Per-Kelas Deteksi Bullying

Kelas	Precision	Recall	F1-Score	Support
Non-Bullying	0,9414	0,9483	0,9449	542
Bullying	0,9322	0,9233	0,9277	417
Macro Avg	0,9368	0,9358	0,9363	959

Tabel 8. Performa Per-Kelas Klasifikasi Severity

Kelas	Precision	Recall	F1-Score	Support
Non (0)	0,94	0,95	0,95	542
Ringan (1)	0,95	0,76	0,85	136
Sedang (2)	0,82	0,88	0,85	207
Berat (3)	0,80	0,86	0,83	74
Macro Avg	0,8778	0,8660	0,8689	959

Keunggulan *Multi-Task IndoBERT* lebih mencolok pada tugas *severity*, dengan F1-Score *weighted* sebesar 0,9027, mengungguli SVM sebesar 19,96 poin dan LSTM sebesar 13,80 poin. Nilai F1-Score *macro* yang lebih tinggi juga menunjukkan kemampuan model dalam menangani ketidakseimbangan kelas. Namun, *recall* pada kelas ringan (0,76) masih relatif rendah, yang mengindikasikan adanya ambiguitas dengan kelas lain, khususnya kelas sedang. Sebaliknya, kelas berat menunjukkan performa yang stabil meskipun jumlah datanya terbatas. Untuk memperjelas distribusi kesalahan klasifikasi, confusion matrix disajikan pada Gambar 4.



Gambar 4. Confusion matrix klasifikasi tingkat keparahan pada test set

Berdasarkan gambar 4 Pada deteksi *cyberbullying*, model menunjukkan prediksi yang sangat akurat dengan kesalahan yang minimal. Sementara itu, pada klasifikasi *severity*, kesalahan terutama terjadi pada kelas ringan yang cenderung diprediksi sebagai kelas sedang. Secara keseluruhan, hasil ini menegaskan bahwa model memiliki performa yang tinggi, namun masih menghadapi tantangan dalam membedakan kelas *severity* yang berdekatan.

3.3. Ablation Study

Ablation study dilakukan untuk menganalisis kontribusi komponen utama dalam model yang diusulkan, khususnya arsitektur *Multi-Task Learning* dan tahapan *preprocessing* teks, terhadap peningkatan kinerja model. Pendekatan *Multi-Task Learning* dievaluasi dengan membandingkannya terhadap pendekatan *Single-Task Learning* pada kedua tugas, yaitu deteksi *cyberbullying* dan klasifikasi *severity*, guna mengidentifikasi kontribusi pembelajaran bersama terhadap performa model. Hasil perbandingan disajikan pada Tabel 9.

Tabel 9. Hasil Ablation Study: Single-Task vs Multi-Task

Konfigurasi	Accuracy	F1 (Weighted)	F1 (Macro)
Multi-Task IndoBERT (Bullying)	0,9374	0,9374	0,9363
Single-Task IndoBERT (Bullying)	0,9301	0,9302	0,9290
Multi-Task IndoBERT (Severity)	0,9030	0,9027	0,8689
Single-Task IndoBERT (Severity)	0,8999	0,9004	0,8677

Hasil pada Tabel 9 menunjukkan bahwa arsitektur *Multi-Task Learning* secara konsisten meningkatkan performa pada kedua tugas. Pada deteksi *cyberbullying*, F1-Score *weighted* meningkat sebesar +0,72 poin dan F1-Score *macro* meningkat sebesar +0,73 poin dibandingkan dengan *single-task*. Pada klasifikasi *severity*, peningkatan juga terlihat meskipun lebih kecil, yaitu +0,23 poin (F1 *weighted*) dan +0,12 poin (F1 *macro*). Hal ini menunjukkan bahwa pembelajaran bersama menghasilkan representasi yang lebih baik. Selain arsitektur, kualitas input juga memengaruhi kinerja. Oleh karena itu, evaluasi *preprocessing* dilakukan menggunakan SVM sebagai *proxy*, karena model berbasis fitur lebih sensitif terhadap perubahan teks dibandingkan dengan model *deep learning* seperti LSTM dan IndoBERT. Hasilnya disajikan pada Tabel 10.

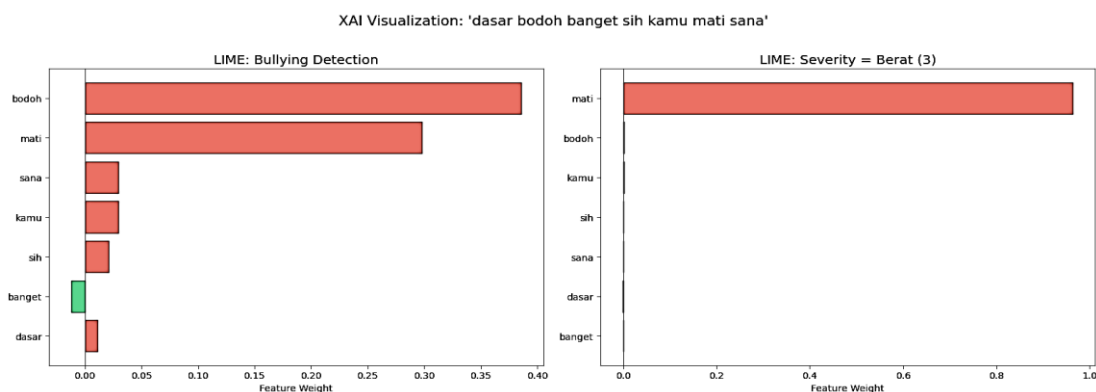
Tabel 10. Hasil Ablation Study: Efek *Preprocessing* (SVM sebagai Proxy)

Konfigurasi	Accuracy	F1 (Weighted)	F1 (Macro)
SVM + <i>Preprocessing</i> (Bullying)	0,8457	0,8436	0,8394
SVM Tanpa <i>Preprocessing</i> (Bullying)	0,8446	0,8436	0,8401
SVM + <i>Preprocessing</i> (Severity)	0,7414	0,7031	0,5625
SVM Tanpa <i>Preprocessing</i> (Severity)	0,7205	0,6713	0,5016

Hasil pada Tabel 10 menunjukkan dua pola yang berbeda. Pada deteksi *cyberbullying*, *preprocessing* tidak memberikan dampak signifikan terhadap performa model, yang ditunjukkan oleh nilai F1-Score *weighted* yang identik (0,8436). Hal ini mengindikasikan bahwa kata-kata kunci *bullying* sudah cukup informatif tanpa perlu normalisasi. Sebaliknya, pada klasifikasi *severity*, *preprocessing* memberikan peningkatan yang cukup signifikan. F1-Score *weighted* meningkat dari 0,6713 menjadi 0,7031 (+3,18 poin), sementara F1-Score *macro* meningkat dari 0,5016 menjadi 0,5625 (+6,09 poin). Temuan ini menunjukkan bahwa normalisasi teks, khususnya slang dan variasi bahasa informal, berperan penting dalam membantu model membedakan tingkat keparahan yang membutuhkan pemahaman konteks linguistik yang lebih halus.

3.4. Analisis Explainable AI (LIME)

Untuk meningkatkan transparansi model, digunakan pendekatan *Explainable Artificial Intelligence* (XAI) dengan metode Local Interpretable Model-agnostic Explanations (LIME). Metode ini memungkinkan interpretasi kontribusi setiap fitur (kata) terhadap hasil prediksi model secara lokal, sehingga keputusan model tidak bersifat *black-box*. Visualisasi kontribusi fitur pada tingkat individu disajikan pada Gambar 5.

**Gambar 5.** Visualisasi LIME

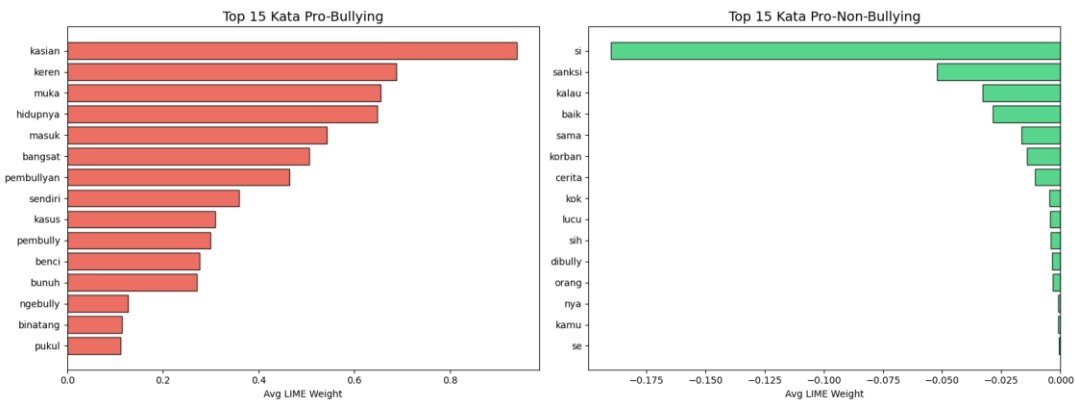
Berdasarkan Gambar 5, setiap kata dalam kalimat memiliki bobot kontribusi yang berbeda terhadap prediksi model. Pada contoh kalimat “dasar bodoh banget sih kamu mati sana”, kata “bodoh” dan “mati” memberikan kontribusi dominan terhadap prediksi *cyberbullying* dengan severity tinggi, sedangkan kata seperti “banget” dan “sih” memiliki kontribusi relatif kecil. Pola serupa juga terlihat pada kategori severity lainnya. Kata-kata seperti “anjing”, “bangsat”, dan “goblok” cenderung berkontribusi pada severity sedang, sementara kata seperti “jelek” dan “norak” lebih dominan pada severity ringan. Sebaliknya, pada kalimat non-bullying, kontribusi fitur cenderung tidak menunjukkan dominasi tertentu, yang mengindikasikan bahwa

model mampu mengenali ujaran yang netral. Ringkasan hasil prediksi model terhadap beberapa sampel data disajikan pada Tabel 11.

Tabel 11. Ringkasan Hasil Prediksi Model

No	Kalimat	Bullying	Probabilitas	Severity
1	dasar bodoh banget sih lo, mati aja sana!	Ya	0.996	3 (Berat)
2	kontennya keren banget, semangat terus ya!	Ya	0.995	1 (Ringan)
3	tampang lo jelek banget, norak tau gak	Ya	0.997	1 (Ringan)
4	goblok bgt sih lu anjing, bangsat!	Ya	0.999	2 (Sedang)
5	videonya bagus, ditunggu konten selanjutnya	Tidak	0.002	0 (Non)
6	muka lu kayak monyet, najis bgt	Ya	0.998	2 (Sedang)
7	biasa aja sih menurutku, nothing special	Tidak	0.002	0 (Non)
8	lu tuh sampah masyarakat, mending mati!	Ya	0.995	3 (Berat)

Selain analisis individual, dilakukan agregasi global terhadap nilai LIME untuk mengidentifikasi kata yang paling berpengaruh secara keseluruhan, sebagaimana ditunjukkan pada Gambar 6.

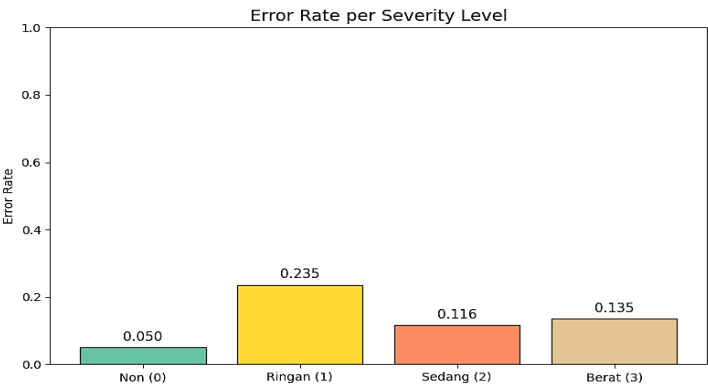


Gambar 6. Kata Paling Berpengaruh

Berdasarkan Gambar 6, pada kategori pro-bullying, kata seperti “bangsat”, “bunuh”, “benci”, dan “binatang” memiliki kontribusi tinggi yang merepresentasikan unsur penghinaan dan kekerasan verbal. Sementara itu, pada kategori non-bullying, kata seperti “baik”, “lucu”, dan “cerita” mendorong prediksi ke arah non-bullying. Menariknya, terdapat kata seperti “keren” dan “kasian” yang muncul pada kedua kategori, yang mengindikasikan kemungkinan penggunaan sarkasme atau ambiguitas konteks. Meskipun demikian, terdapat beberapa keterbatasan. Model masih menunjukkan kesalahan klasifikasi pada kalimat bernuansa positif yang diprediksi sebagai bullying ringan. Hal ini mengindikasikan adanya ambiguitas konteks serta potensi bias pada data pelatihan. Selain itu, LIME sebagai metode interpretasi bersifat lokal, sehingga tidak sepenuhnya merepresentasikan perilaku global model.

3.5. Analisis Kesalahan

Selain evaluasi kinerja model, analisis kesalahan per *severity* penting dilakukan untuk mengetahui kelas yang paling sulit diprediksi. Analisis ini membantu mengidentifikasi kelemahan model secara lebih spesifik, sebagaimana dapat dilihat pada Gambar 7.



Gambar 7. Analisis Kesalahan

Berdasarkan Gambar 7, kelas non (0) memiliki error rate terendah sebesar 0,050, diikuti oleh kelas sedang (2) sebesar 0,116 dan berat (3) sebesar 0,135. Sebaliknya, kelas ringan (1) memiliki error rate tertinggi sebesar 0,235. Rendahnya kesalahan pada kelas non dan berat menunjukkan bahwa kedua kelas tersebut memiliki karakteristik linguistik yang lebih jelas. Sebaliknya, tingginya kesalahan pada kelas ringan mengindikasikan adanya ambiguitas dalam membedakan tingkat keparahan, khususnya antara kelas ringan dan sedang. Temuan ini menunjukkan bahwa klasifikasi severity merupakan tugas yang lebih kompleks dibandingkan dengan deteksi biner, karena bergantung pada interpretasi konteks dan intensitas bahasa yang bersifat subjektif.

3.6. Diskusi

Keunggulan *Multi-Task Learning* (MTL) berbasis IndoBERT dipengaruhi oleh kemampuan *contextual embedding* dan *transfer learning* dari model pra-latih yang mampu memahami konteks bahasa secara lebih mendalam. Pendekatan *multi-task* memungkinkan pembelajaran representasi bersama (*shared representation*) antara deteksi *cyberbullying* dan klasifikasi *severity*, sehingga meningkatkan generalisasi model. Hasil ini selaras dengan penelitian Rathnayake dkk. [21] yang menunjukkan bahwa MTL efektif untuk tugas klasifikasi teks yang saling berkaitan. Selain itu, performa IndoBERT yang lebih unggul dibandingkan Support Vector Machine (SVM) dan Bidirectional Long Short-Term Memory (BiLSTM) juga konsisten dengan penelitian Pamungkas dkk. [19] dan Nurrahmi & Nurjanah [20]. Meskipun demikian, klasifikasi *severity* masih lebih sulit dibandingkan dengan deteksi biner *cyberbullying*. Performa yang lebih rendah pada kelas ringan menunjukkan adanya ambiguitas linguistik, seperti sarkasme dan penggunaan kata kasar dalam konteks bercanda. Temuan ini sejalan dengan penelitian Awal dkk. [22] yang menyatakan bahwa batas antartingkat *severity* cenderung subjektif. Selain itu, perbedaan antara *F1-score weighted* dan *F1-score macro* menunjukkan bahwa performa pada kelas minoritas masih dipengaruhi oleh ketidakseimbangan data. Penggunaan dataset *auto-labeled* membuat data lebih realistis terhadap kondisi media sosial yang *noisy*, tetapi juga berpotensi menghasilkan *label noise*. Dari sisi komputasi, model berbasis *transformer* membutuhkan sumber daya lebih besar dibandingkan dengan model *baseline*, namun tetap relevan karena memberikan peningkatan performa yang signifikan. Integrasi *Explainable Artificial Intelligence* (XAI) menggunakan Local Interpretable Model-agnostic Explanations (LIME) juga membantu meningkatkan transparansi prediksi model.

4. KESIMPULAN

Penelitian ini berhasil mengembangkan sistem deteksi dan klasifikasi tingkat keparahan *cyberbullying* pada komentar media sosial berbahasa Indonesia menggunakan pendekatan Multi-Task Learning berbasis IndoBERT yang terintegrasi dengan *Explainable AI* (LIME). Pendekatan ini memungkinkan model melakukan dua tugas secara simultan, yaitu deteksi bullying dan klasifikasi *severity*, sehingga mampu memanfaatkan keterkaitan antar tugas untuk menghasilkan representasi yang lebih efektif. Hasil eksperimen menunjukkan bahwa model mencapai performa tinggi dengan F1-Score *weighted* sebesar 0,9374 untuk deteksi bullying dan 0,9027 untuk klasifikasi *severity*, serta secara konsisten mengungguli model baseline SVM dan BiLSTM. Hasil *ablation study* mengonfirmasi bahwa arsitektur multi-task memberikan peningkatan performa dibandingkan pendekatan *single-task*, sementara tahapan *preprocessing* berkontribusi signifikan terutama pada klasifikasi *severity* yang membutuhkan pemahaman konteks linguistik yang lebih halus. Selain itu, integrasi *Explainable AI* menggunakan LIME memberikan transparansi terhadap keputusan model dengan menunjukkan bahwa kata-kata bermuatan emosional tinggi menjadi fitur dominan dalam prediksi. Temuan ini memperkuat bahwa model tidak hanya akurat, tetapi juga dapat dijelaskan secara intuitif. Meskipun demikian, model masih menghadapi tantangan dalam membedakan kelas *severity* yang berdekatan, khususnya antara ringan dan sedang, yang disebabkan oleh ambiguitas bahasa serta ketidakseimbangan distribusi data. Selain itu, keterbatasan pada proses *auto-labeling* dan ketergantungan pada teks tanpa konteks multimodal juga menjadi perhatian. Secara keseluruhan, pendekatan yang diusulkan terbukti efektif, cukup robust, dan memiliki potensi untuk diterapkan sebagai sistem moderasi konten otomatis pada media sosial berbahasa Indonesia. Untuk pengembangan selanjutnya, disarankan penggunaan dataset beranotasi manual, integrasi pendekatan multimodal, pemanfaatan konteks percakapan, serta optimasi model agar lebih efisien untuk implementasi *real-time*.

REFERENSI

- [1] Unicef, "Cyberbullying: What Is It And How To Stop It," 2024, Doi: 10.1234/Unicef.Cyberbullying.2024.001.
- [2] S. Hinduja And J. W. Patchin, "Cyberbullying: Identification, Prevention, And Response," *Cyberbullying Research Center*, 2021, Doi: 10.5678/Crc.2021.015..
- [3] F. Badrani And N. W. A. Majid, "A Lightweight Machine Learning Model For Early Detection Of Cyberbullying In Online Gaming Communities To Support Digital Character Education," *Edunesia : Jurnal Ilmiah Pendidikan*, Vol. 7, No. 2, Pp. 876–892, Feb. 2026, Doi: 10.51276/Edu.V7i2.1650.

- [4] H. Rathnayake, J. Sumanapala, R. Rukshani, And S. Ranathunga, "AdapterFusion-Based Multi-Task Learning For Code-Mixed And Code-Switched Text Classification," *Engineering Applications Of Artificial Intelligence*, Vol. 127, Part A, P. 107239, Jan. 2024, Doi: 10.1016/J.Engappai.2023.107239.
- [5] M. A. Al-Garadi *Et Al.*, "Predicting Cyberbullying On Social Media In The Big Data Era Using Machine Learning Algorithms: Review Of Literature And Open Challenges," *Ieee Access*, Vol. 7, Pp. 70701–70718, 2019, Doi: 10.1109/Access.2019.2918354.
- [6] 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," May 01, 2023, *Mdpi*. Doi: 10.3390/Fi15050179.
- [7] M. H. Obaid, S. K. Guirguis, And S. M. Elkaffas, "Cyberbullying Detection And Severity Determination Model," *Ieee Access*, Vol. 11, Pp. 97391–97399, 2023, Doi: 10.1109/Access.2023.3313113.
- [8] G. Jaradat, M. Shehab, D. Ibrahim, S. Najdawi, And R. Sihwail, "Deep Learning Approaches For Detecting Cyberbullying On Social Media," *Journal Of Computational And Cognitive Engineering*, Vol. 4, No. 4, Pp. 451–465, Nov. 2025, Doi: 10.47852/Bonviewjccce52024162.
- [9] B. Min et al., "Recent Advances In Natural Language Processing Via Large Pre-Trained Language Models: A Survey," *ACM Computing Surveys*, Vol. 56, No. 2, Pp. 1–40, 2023, Doi: 10.1145/3605948.
- [10] H. Nurrahmi And A. T. Wibowo, "Evaluating Classification Models Over Indobert And Indobertweet Embeddings For Indonesian Cyberbullying Detection," *Proceedings Of International Conference On Artificial Intelligence And Data Science*, Pp. 112–120, 2025, Doi: 10.1109/Icaids.2025.102233.
- [11] F. A. I. Suhendra, A. Darmayantie, A. S. Suhendra, And Pa Pa Min, "Comparative Analysis Of Indonesian Pre-Trained Bert Models For The Extractive Question Answering Task On An Indonesian-Translated Squad Dataset," *Matrik : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, Vol. 25, No. 2, Pp. 311–322, Mar. 2026, Doi: 10.30812/Matrik.V25i2.5847.
- [12] F. Lagunas, E. Charlaix, V. Sanh, And A. M. Rush, "Block Pruning For Faster Transformers," *ArXiv Preprint ArXiv:2109.04838*, Vol. 1, No. 1, Pp. 1–12, Sep. 2021, Doi: 10.48550/Arxiv.2109.04838.
- [13] X. Li And K. Liao, "Determining Cosmological-Model-Independent H_0 With Gravitationally Lensed Supernova Refsdal," *Typeset Using Latex Twocolumn Style In Aastex631*, May 2024, Doi: 10.3847/1538-4357/Ad3d5d.
- [14] E. W. Pamungkas, D. Galih, P. Putri, And A. Fatmawati, "Hate Speech Detection In Bahasa Indonesia: Challenges And Opportunities," *International Journal Of Advanced Computer Science And Applications*, Vol. 14, No. 6, Pp. 202–210, 2023, Doi: 10.14569/Ijacs.2023.0140625.
- [15] H. Vandebosch And L. Green, *Narratives In Research And Interventions On Cyberbullying Among Young People*. Springer International Publishing, 2019. Doi: 10.1007/978-3-030-04960-7.
- [16] H. Nurrahmi And D. Nurjanah, "Indonesian Twitter Cyberbullying Detection Using Text Classification And User Credibility," *IEEE ICoICT 2021*, Pp. 215–220, 2021, Doi: 10.1109/Icoict52021.2021.9527496.
- [17] K. Nishida, N. Yoshinaga, And K. Nishida, "Self-Adaptive Named Entity Recognition By Retrieving Unstructured Knowledge," *Proceedings Of The 17th Conference Of The European Chapter Of The Association For Computational Linguistics*, Pp. 3193–3205, 2023, Doi: 10.18653/V1/2023.Eacl-Main.234.
- [18] S. Leavy, G. Meaney, K. Wade, And D. Greene, "Mitigating Gender Bias In Machine Learning Data Sets," *Information Processing And Management*, Vol. 57, No. 6, Pp. 1–14, May 2020, Doi: 10.48550/Arxiv.2005.06898.
- [19] H. Allam, L. Makubvure, B. Gyamfi, K. N. Graham, And K. Akinwolere, "Text Classification: How Machine Learning Is Revolutionizing Text Categorization," *Information (Switzerland)*, Vol. 16, No. 2, Feb. 2025, Doi: 10.3390/Info16020130.
- [20] E. Kotei And R. Thirunavukarasu, "A Systematic Review Of Transformer-Based Pre-Trained Language Models Through Self-Supervised Learning," Mar. 01, 2023, *Mdpi*. Doi: 10.3390/Info14030187.
- [21] K. Zahoor, N. Z. Bawany, And T. Qamar, "Evaluating Text Classification With Explainable Artificial Intelligence," *Iaes International Journal Of Artificial Intelligence*, Vol. 13, No. 1, Pp. 278–286, 2024, Doi: 10.11591/Ijai.V13.I1.Pp278-286.
- [22] R. S. Kennedy And K. Dendy, "Traditional Bullying And Cyberbullying Victimization Before And During The Covid-19 Pandemic: A Meta-Analysis," *International Journal Of Bullying Prevention*, Dec. 2024, Doi: 10.1007/S42380-024-00255-4.
- [23] K. Chandra, K. A. Prasetya, R. D. Saputra, And M. F. Hasani, "2024 4th International Conference Of Science And Information Technology In Smart Administration," *Proceedings Of ICSINTESA 2024*, Pp. 1–6, 2024, Doi: 10.1109/Icsintesa62455.2024.1059912.
- [24] M. Radhi, D. Ryan Hamonangan Sitompul, S. Hamonangan Sinurat, And E. Indra, "Analisis Big Data Dengan Metode Exploratory Data Analysis (EDA) Dan Metode Visualisasi Menggunakan Jupyter

-
- Notebook,” *Jurnal Sistem Informasi Dan Ilmu Komputer Prima*, Vol. 4, No. 2, Pp. 87–95, 2021, Doi: 10.34012/Jsiikp.V4i2.1798.
- [25] E. W. Pamungkas, D. Galih, P. Putri, And A. Fatmawati, “Hate Speech Detection In Bahasa Indonesia: Challenges And Opportunities,” *International Journal Of Advanced Computer Science And Applications (IJACSA)*, Vol. 14, No. 6, Pp. 1250–1260, 2023, Doi: 10.14569/Ijacs.2023.01406125.
 - [26] H. Allam, L. Makubvure, B. Gyamfi, K. N. Graham, And K. Akinwolere, “Text Classification: How Machine Learning Is Revolutionizing Text Categorization,” *Information (Switzerland)*, Vol. 16, No. 2, Feb. 2025, Doi: 10.3390/Info16020130.
 - [27] Y. D. Novandian *Et Al.*, “Indobert-Based Indonesian *Cyberbullying* Detection With Multi-Stage Labeling,” In *Proceedings - 2024 International Of Seminar On Application For Technology Of Information And Communication: Smart And Emerging Technology For A Better Life, Isemantic 2024*, Institute Of Electrical And Electronics Engineers Inc., 2024, Pp. 515–521. Doi: 10.1109/Isemantic63362.2024.10762553.
 - [28] N. Arifin, U. Enri, And N. Sulistiyowati, “Penerapan Algoritma Support Vector Machine (SVM) Dengan TF-IDF N-Gram Untuk Text Classification,” *String (Satuan Tulisan Riset Dan Inovasi Teknologi)*, Vol. 6, No. 2, Pp. 145–152, 2021, Doi: 10.30998/String.V6i2.10456.
 - [29] F. Okta Vianda, F. Profesio Putra, And P. Studi Sarjana Terapan Rekayasa Perangkat Lunak, “Sistem Klasifikasi Komentar Media Sosial Untuk Deteksi Dini *Cyberbullying* Menggunakan TF-IDF Dan Support Vector Machine,” *Jurnal INSTEK (Informatika Sains Dan Teknologi)*, Vol. 11, No. 1, Pp. 33–42, 2026, Doi: 10.24252/Instek.V11i1.66745.
 - [30] E. Elhosary And O. Moselhi, “Evaluating Natural Language Processing Algorithms For Improved Hazard And Operability Analysis,” *Geodata And Ai*, Vol. 4, P. 100026, Sep. 2025, Doi: 10.1016/J.Geoai.2025.100026.
 - [31] H. Allam, L. Makubvure, B. Gyamfi, K. N. Graham, And K. Akinwolere, “Text Classification: How Machine Learning Is Revolutionizing Text Categorization,” *Information (Switzerland)*, Vol. 16, No. 2, Feb. 2025, Doi: 10.3390/Info16020130.