JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 3, September 2025, Pp. 2666-2678

DETEKSI CYBERBULLYING MULTIKELAS BERKINERJA TINGGI: ENSEMBLE ROBERTA-LARGE DENGAN PRESISI CAMPURAN

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

1. Teknologi Informasi, Fakultas Sains dan Teknologi, Universitas Islam Negeri Walisongo Semarang, Indonesia
2. Teknologi Informasi, Fakultas Sains dan Teknologi, Universitas Islam Negeri Walisongo Semarang, Indonesia
3. Teknologi Informasi, Fakultas Sains dan Teknologi, Universitas Islam Negeri Walisongo Semarang, Indonesia
4. Teknologi Informasi, Fakultas Sains dan Teknologi, Universitas Islam Negeri Walisongo Semarang, Indonesia

Article Info
Kata Kunci: *Cyberbullying*, *Deep Learning*, *Ensemble Learning*, Klasifikasi Teks, Presisi Campuran, RoBERTa

Keywords: *Cyberbullying, Deep Learning, Ensemble Learning, Mixed Precision, RoBERTa, Text Classification*

Article history:
Received 25 May 2025
Revised 10 June 2025
Accepted 20 June 2025
Available online 1 September 2025

DOI :
https://doi.org/10.29100/jipi.v10i3.8056

* Corresponding author.
Khothibul Umam
E-mail address:
khothibul_umam@walisongo.ac.id

ABSTRAK

Isu *cyberbullying* yang terus berkembang di lingkungan digital telah menjadi perhatian global serius, menimbulkan dampak negatif signifikan dan menyoroti kebutuhan mendesak akan sistem deteksi otomatis. Tujuan primer penelitian ini adalah mengembangkan dan mengevaluasi sistem klasifikasi *cyberbullying* multikelas yang efektif, mampu mengidentifikasi kelas-kelas *age, ethnicity, gender,* dan *religion*, sekaligus membedakannya dari konten *not_cyberbullying* dan *other_cyberbullying*. Desain penelitian ini adalah eksperimental, berfokus pada *fine-tuning* model bahasa besar untuk tugas klasifikasi teks. Metodologi yang diterapkan melibatkan *fine-tuning* model RoBERTa-Large menggunakan dataset terlabel multikelas sebanyak 47.692 tweet. Untuk meningkatkan robustisitas dan generalisasi model, digunakan teknik *ensemble learning* melalui *soft voting* dari tiga model RoBERTa-Large yang dilatih dengan *seed* yang berbeda. Pelatihan dilakukan dengan presisi campuran (FP16) untuk efisiensi komputasi. Hasil utama menunjukkan bahwa model ensemble ini mencapai kinerja yang solid dan kompetitif pada *test set* untuk deteksi *cyberbullying* multikelas, dengan Akurasi 0.87 dan F1-Score (Weighted) sebesar 0.86. Model menunjukkan kinerja yang sangat baik pada kelas-kelas *age, ethnicity, gender,* dan *religion* tersebut, namun masih menghadapi tantangan pada klasifikasi kelas *not_cyberbullying* dan *other_cyberbullying*. Kesimpulannya, sistem ini membuktikan efektivitas signifikan dari RoBERTa-Large dalam konfigurasi ensemble untuk deteksi *cyberbullying* multikelas, menunjukkan kemampuan deteksi yang kuat secara keseluruhan dan sangat baik pada kategori-kategori tertentu, memberikan dasar kuat untuk aplikasi pencegahan *cyberbullying* di dunia nyata.

ABSTRACT

The escalating issue of cyberbullying in the digital environment has become a serious global concern, leading to significant negative impacts and highlighting the urgent need for automatic detection systems. The primary objective of this research is to develop and evaluate an effective multiclass cyberbullying classification system, capable of identifying the age, ethnicity, gender, and religion classes, while simultaneously distinguishing them from not_cyberbullying and other_cyberbullying content. The research design is experimental, focusing on fine-tuning large language models for text classification tasks. The implemented methodology involves fine-tuning a RoBERTa-Large model using a multiclass labeled dataset of 47,692 tweets. To enhance model robustness and generalization, an ensemble learning technique using soft voting from three RoBERTa-Large models, each trained with a different seed, was employed. Training was conducted with mixed precision (FP16) for computational efficiency. The main outcomes show that this ensemble model achieved solid and competitive performance on the test set for multiclass cyberbullying detection, with an Accuracy of 0.87 and a Weighted F1-Score of 0.86. The model demonstrated excellent performance on these age, ethnicity, gender, and religion classes, but still faced challenges in classifying not_cyberbullying and other_cyberbullying classes. In conclusion, this system proves the significant effectiveness of RoBERTa-Large in an

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 3, September 2025, Pp. 2666-2678

ensemble configuration for multiclass cyberbullying detection, demonstrating strong overall detection capabilities and excellent performance on specific categories, providing a strong foundation for real-world cyberbullying prevention applications.

.

I. PENDAHULUAN

FENOMENA *cyberbullying* telah muncul sebagai ancaman siber yang signifikan di era digital, dengan dampak luas terhadap individu dan masyarakat [1]. Berbeda dari bentuk bullying tradisional, cyberbullying memanfaatkan platform komunikasi elektronik, memungkinkan tindakan agresif yang disengaja dan berulang secara anonim dan dengan jangkauan yang sangat luas [1-2]. Dampak negatif dari cyberbullying sangat serius, seringkali menyebabkan gangguan kesehatan mental yang parah seperti depresi, kecemasan, isolasi sosial, dan bahkan memicu kasus bunuh diri pada korban [1]. Mengingat miliaran pengguna aktif di berbagai platform media sosial dan volume data daring yang terus meningkat, deteksi cyberbullying secara manual menjadi tidak praktis dan tidak efisien [3-4]. Oleh karena itu, pengembangan sistem deteksi cyberbullying otomatis yang efektif menjadi sangat krusial untuk menciptakan lingkungan daring yang lebih aman dan mendukung kesejahteraan digital.

Meskipun urgensi deteksi *cyberbullying* telah mendorong banyak penelitian, tugas ini dihadapkan pada sejumlah tantangan kompleks, terutama dalam konteks klasifikasi multikelas. Sifat bahasa di platform media sosial yang seringkali ambigu, melibatkan sarkasme, ironi, dan bahasa *slang*, menyulitkan model untuk menangkap niat tersembunyi dari suatu ujaran [3],[5]. Selain itu, *cyberbullying* dapat bermanifestasi dalam berbagai bentuk spesifik, seperti yang terkait dengan usia, etnis, gender, atau agama, yang masing-masing mungkin memiliki pola linguistik yang berbeda dan memerlukan pemahaman kontekstual yang mendalam untuk diidentifikasi secara akurat [4], [6]. Tantangan juga muncul dari sifat dinamis *cyberbullying* yang terus berevolusi, di mana para pelaku secara konsisten mengembangkan taktik baru untuk menghindari deteksi [3]. Lebih lanjut, tren *cyberbullying* menunjukkan evolusi yang konstan, di mana pelaku seringkali menggunakan bahasa kode, emoji dengan makna ganda, atau bentuk-bentuk ujaran implisit lainnya untuk menghindari deteksi oleh sistem keamanan standar. Peningkatan kompleksitas pola ujaran berbahaya ini menuntut pengembangan sistem deteksi yang tidak hanya akurat dalam mengenali pola-pola yang sudah dikenal, tetapi juga adaptif dan mampu menangkap nuansa linguistik yang lebih halus dan terus berubah. Oleh karena itu, terdapat kebutuhan yang jelas untuk mengembangkan model yang lebih canggih dan *robust* yang mampu mengatasi kompleksitas linguistik dan keragaman kategori dalam deteksi *cyberbullying* multikelas.

Kemajuan terkini dalam *Natural Language Processing* (NLP), khususnya dengan munculnya Model Bahasa Besar (LLM) berbasis *transformer* seperti RoBERTa, menawarkan solusi menjanjikan untuk mengatasi tantangan tersebut [5]. Model-model ini, yang dilatih pada korpus teks yang sangat besar, mampu menghasilkan representasi kontekstual yang kaya dan menangkap dependensi jangka panjang dalam teks, menjadikannya sangat cocok untuk tugas klasifikasi teks yang kompleks seperti deteksi *cyberbullying* [5]. Secara khusus, kemampuan RoBERTa-Large dalam memahami konteks dua arah secara mendalam memungkinkannya untuk lebih baik dalam menginterpretasi nuansa bahasa seperti sarkasme atau ironi, yang sering menjadi tantangan bagi model *deep learning* generasi sebelumnya seperti CNN atau LSTM yang mungkin tidak menangkap dependensi kontekstual jarak jauh seefektif arsitektur *Transformer*. Dalam penelitian ini, kami memanfaatkan kekuatan RoBERTa-Large yang telah terbukti unggul dalam berbagai tugas NLP. Untuk lebih meningkatkan *robustness*, akurasi, dan kemampuan generalisasi model, kami mengusulkan dan mengimplementasikan teknik *ensemble learning* melalui *soft voting* dari beberapa model RoBERTa-Large yang dilatih secara independen [7]. Pendekatan *ensemble* ini tidak hanya bertujuan meningkatkan akurasi secara umum, tetapi secara spesifik dirancang untuk meningkatkan ketahanan (*robustness*) terhadap variasi linguistik dan ambiguitas dalam ujaran *cyberbullying*. Dengan menggabungkan prediksi dari beberapa model RoBERTa-Large yang memiliki 'perspektif' sedikit berbeda (karena dilatih dengan *seed* acak yang berbeda), sistem *ensemble* dapat mengurangi risiko kesalahan interpretasi tunggal terhadap teks yang ambigu atau mengandung bahasa *slang* yang kompleks, sebuah keunggulan dibandingkan jika hanya mengandalkan satu model *deep learning* tunggal, secanggih apapun arsitekturnya. Kontribusi utama dari penelitian ini adalah pengembangan dan evaluasi sistem klasifikasi *cyberbullying* multikelas berkinerja tinggi. Sistem yang kami usulkan, berbasis *ensemble* RoBERTa-Large, didemonstrasikan mampu mengidentifikasi beragam jenis *cyberbullying* dengan akurasi dan stabilitas yang signifikan, menawarkan fondasi kuat untuk aplikasi

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 3, September 2025, Pp. 2666-2678

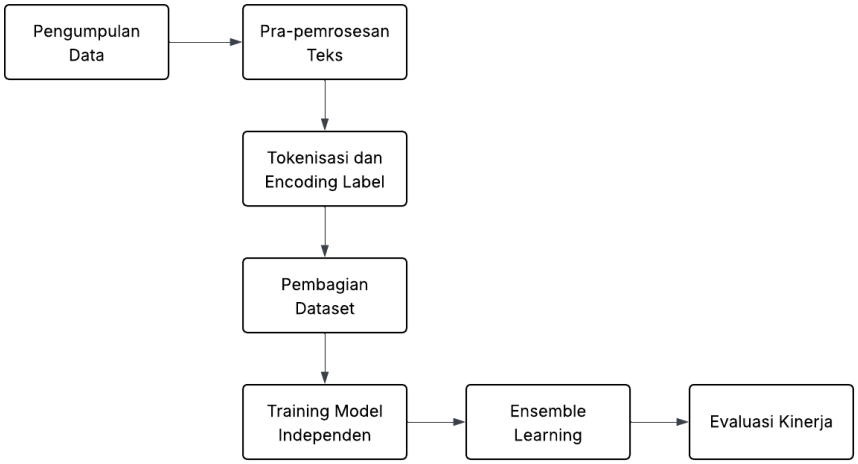
pencegahan *cyberbullying* di dunia nyata.

II. METODE PENELITIAN

A. Arsitektur Sistem

Penelitian ini mengimplementasikan pendekatan berbasis *ensemble deep learning* untuk deteksi *cyberbullying* multikelas pada data media sosial. Seluruh implementasi model dan eksperimen dilakukan menggunakan bahasa pemrograman Python dengan memanfaatkan pustaka *deep learning* PyTorch dan Hugging Face Transformers pada platform komputasi akselerasi GPU (misalnya, Google Colaboratory) untuk memfasilitasi pelatihan model yang intensif secara komputasi. Alur penelitian dirancang untuk secara sistematis memproses data mentah, melatih model-model pembelajaran mendalam, dan menggabungkan hasilnya untuk mendapatkan prediksi yang robust. Metodologi klasifikasi teks secara umum telah mengalami evolusi signifikan, beralih dari pendekatan tradisional ke pembelajaran mendalam yang lebih canggih [8].

Secara umum, alur kerja yang diterapkan dalam penelitian ini terdiri dari beberapa tahapan utama, sebagaimana diilustrasikan pada Gambar 1. Tahapan-tahapan tersebut meliputi: (1) Pengumpulan dan Pemahaman Data, di mana data tweet awal diperoleh dari platform yang tersedia secara publik seperti Kaggle dan karakteristiknya dianalisis; (2) Pra-pemrosesan Teks, yang melibatkan serangkaian langkah untuk membersihkan dan menstandarisasi data teks; (3) Tokenisasi dan Encoding Label, untuk mengubah teks dan label menjadi format yang dapat diproses oleh model; (4) Pembagian Dataset, membagi data menjadi set pelatihan, validasi, dan pengujian; (5) Pelatihan Model RoBERTa-Large Independen, di mana beberapa model RoBERTa-Large dilatih secara terpisah dengan inisialisasi yang berbeda; (6) Ensemble Learning, yang mengintegrasikan prediksi dari model-model yang dilatih secara independen menggunakan strategi *soft voting* [9]; dan (7) Evaluasi Kinerja, di mana model ensemble dinilai menggunakan metrik kuantitatif dan visualisasi pada *test set* yang tidak terlihat.



Gambar. 1. Alur Kerja Sistem Deteksi *Cyberbullying* Multikelas

B. Dataset

Dataset yang digunakan dalam penelitian ini adalah koleksi *tweet* berbahasa Inggris yang diperoleh dari platform Kaggle. Pemilihan dataset spesifik ini didasarkan pada beberapa pertimbangan krusial. Pertama, dataset ini memiliki volume data yang substansial, yaitu 47.692 *tweet* unik, yang sangat memadai untuk kebutuhan pelatihan dan evaluasi model pembelajaran mendalam (*deep learning*) seperti RoBERTa-Large. Kedua, dataset ini menyediakan anotasi kelas *cyberbullying* yang rinci, mencakup enam kategori berbeda (yaitu: *age, ethnicity, gender, religion, not_cyberbullying,* dan *other_cyberbullying*), yang secara langsung mendukung tujuan penelitian ini untuk mengembangkan sistem klasifikasi multikelas. Ketiga, dataset ini memiliki skor *usability* sempurna (10.00) di platform Kaggle, mengindikasikan kualitas, kelengkapan, dan kemudahan penggunaan yang tinggi menurut penilaian komunitas, yang meminimalkan potensi masalah terkait kualitas data mentah. Selain itu, ketersediaan dataset secara publik juga memastikan transparansi dan memungkinkan replikabilitas penelitian oleh peneliti lain di masa mendatang.

Karakteristik data media sosial, khususnya *tweet*, seringkali menunjukkan penggunaan bahasa informal, seperti singkatan, ejaan tidak standar, *emoticon*, dan struktur kalimat yang longgar [3]. Aspek-aspek ini menjadi pertimbangan penting dalam merancang tahapan pra-pemrosesan teks untuk memastikan bahwa model dapat menangani variasi bahasa tersebut secara efektif. Untuk memberikan pemahaman yang komprehensif mengenai karakteristik dataset, berikut adalah analisis lebih lanjut yang mencakup distribusi kelas dan visualisasi data:

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 3, September 2025, Pp. 2666-2678

1. Distribusi Kelas

Distribusi *tweet* di seluruh kategori kelas disajikan dalam Tabel I. Dengan jumlah total *tweet* 47.692, dataset ini menunjukkan distribusi kelas yang relatif seimbang (lihat Tabel I), sebuah keuntungan mengingat ketidakseimbangan kelas (*class imbalance*) merupakan tantangan umum dalam banyak dataset deteksi *cyberbullying* yang dapat memengaruhi evaluasi model [4]. Keseimbangan pada dataset ini membantu dalam pelatihan dan evaluasi model yang lebih objektif untuk setiap kelas. Karakteristik keseimbangan relatif ini menjadi dasar pertimbangan dalam strategi pembagian dataset untuk mempertahankan representasi kelas yang adil.

TABEL I
DISTRIBUSI KATEGORI CYBERBULLYING DALAM DATASET

Cyberbullying_type	Jumlah Tweet
Religion	7998
Age	7992
Gender	7973
Ethnicity	7961
Not_cybeybullying	7945
Other_cybeybullying	7823

2. Visualisasi Dataset

Untuk memberikan gambaran yang lebih jelas mengenai komposisi dataset dan karakteristik verbal tiap kelas, analisis *word cloud* untuk setiap kelas juga disajikan pada Gambar 2. Ini memberikan wawasan awal tentang kata-kata kunci atau frasa yang dominan yang terkait dengan setiap jenis *cyberbullying*.

Gambar. 2. Word cloud untuk setiap kategori cyberbullying

C. Pra-pemrosesan Teks

Tahap pra-pemrosesan teks adalah langkah krusial dalam *Natural Language Processing* (NLP) untuk membersihkan dan menstandarisasi data teks, sehingga model dapat memprosesnya dengan lebih efektif dan mengurangi kebisingan (*noise*) yang tidak perlu [10]. Dalam konteks data media sosial seperti *tweet*, yang seringkali tidak terstruktur dan mengandung banyak elemen informal, pra-pemrosesan menjadi semakin penting [3]. Namun, untuk model berbasis Transformer seperti RoBERTa, pra-pemrosesan yang terlalu agresif dapat menghilangkan informasi kontekstual penting yang telah dipelajari oleh model selama tahap *pre-training* [5]. Oleh karena itu, pendekatan pra-pemrosesan minimalis diterapkan dalam penelitian ini. Langkah-langkah pra-pemrosesan yang dilakukan meliputi:

1. Penghapusan Tautan (*URL Removal*): Semua *URL* (misalnya, http://t.co/xyz atau https://...) dihapus dari *tweet* karena tautan tersebut tidak memberikan informasi tekstual yang relevan untuk klasifikasi *cyberbullying* dan dapat dianggap sebagai *noise*.
2. Penghapusan Sebutan Pengguna (*User Mention Removal*): Sebutan pengguna (misalnya, *@username*) dihapus dari *tweet*. Meskipun sebutan pengguna mengindikasikan interaksi antar pengguna, nama pengguna itu sendiri tidak secara langsung berkontribusi pada penentuan kategori *cyberbullying* dari konten teks.
3. Normalisasi Teks (*Text Normalization*): Teks dinormalisasi untuk menangani inkonsistensi seperti ejaan tidak standar atau penggunaan huruf besar/kecil yang tidak konsisten. Ini mencakup konversi semua teks menjadi

huruf kecil (*lowercasing*). Meskipun normalisasi secara umum dapat mencakup *stemming* atau *lemmatization*, pendekatan minimalis untuk transformer seringkali menghindari langkah-langkah ini untuk mempertahankan bentuk kata asli. [11] juga membahas pentingnya normalisasi teks untuk bahasa tertentu.

4. Penghapusan Karakter Non-Alfabetik dan Numerik: Karakter atau simbol yang bukan huruf alfabet atau angka, yang tidak memberikan informasi semantik, dihapus. Ini membantu membersihkan teks dari simbol-simbol yang tidak relevan.

Langkah-langkah ini memastikan bahwa teks yang masuk ke model Transformer bersih dari *noise* yang jelas sambil mempertahankan sebagian besar informasi kontekstual yang diperlukan oleh model *pre-trained*.

D. Tokenisasi Teks dan Encoding Label

Setelah tahap pra-pemrosesan, teks perlu diubah menjadi representasi numerik yang dapat dipahami oleh model pembelajaran mendalam. Tahap ini melibatkan dua komponen utama: tokenisasi teks dan encoding label.

1. Tokenisasi Teks**:**

Tokenisasi adalah proses memecah urutan teks menjadi unit-unit yang lebih kecil, yang disebut token. Untuk model berbasis Transformer seperti RoBERTa-Large, proses tokenisasi sangat spesifik dan menggunakan *tokenizer* yang telah dilatih sebelumnya (pre-trained tokenizer) yang sesuai dengan arsitektur model. RoBERTa, sebagai varian dari BERT, menggunakan tokenisasi *Byte Pair Encoding (BPE)* atau *WordPiece* yang menangani kata-kata yang tidak dikenal dengan memecahnya menjadi sub-kata (*subword units*) yang lebih kecil [12]. Pendekatan ini memungkinkan model untuk menangani kosakata yang besar dan kata-kata di luar kosakata (*Out-Of-Vocabulary/OOV*) secara efektif, tanpa kehilangan informasi semantik yang signifikan

Setiap *tweet* akan diubah menjadi urutan token ID numerik. Selain itu, *tokenizer* juga menambahkan token khusus yang penting untuk operasi model Transformer:

- [CLS] (Classifier): Token khusus yang ditambahkan di awal setiap urutan. Vektor representasi dari token ini pada lapisan keluaran terakhir seringkali digunakan sebagai representasi agregat dari seluruh urutan untuk tugas klasifikasi [5].
- [SEP] (Separator): Token khusus yang digunakan untuk memisahkan segmen teks (meskipun dalam kasus klasifikasi teks tunggal, ini sering ditempatkan di akhir kalimat).
- *Attention Mask*: Sebuah *array* yang menunjukkan token mana yang harus diperhatikan oleh model (nilai 1 untuk token nyata, 0 untuk token *padding*). Ini mencegah model memperhatikan token *padding* yang tidak relevan.
- *Token Type IDs* (Segment IDs): *Array* yang mengidentifikasi segmen tempat token berada (misalnya, 0 untuk kalimat pertama, 1 untuk kalimat kedua). Dalam kasus teks tunggal, ini biasanya hanya berisi 0.

Panjang maksimum urutan token ditentukan sebesar 128 untuk memastikan konsistensi *input* ke model dan untuk mengelola memori komputasi. Setiap urutan yang lebih pendek dari panjang maksimum akan diisi dengan token *padding* hingga mencapai panjang maksimum, dan setiap urutan yang lebih panjang akan dipotong.

2. Encoding Label (*One-Hot Encoding*):

Label kategori *cyberbullying* (yaitu, *age*, *ethnicity*, *gender*, *religion*, *not_cyberbullying*, *other_cyberbullying*) adalah data kategorikal. Untuk melatih model klasifikasi, label-label ini perlu diubah menjadi format numerik. Pendekatan yang umum digunakan adalah *One-Hot Encoding*.

Dalam *One-Hot Encoding*, setiap kategori unik diubah menjadi vektor biner di mana hanya satu elemen yang bernilai 1 (menunjukkan kategori yang aktif) dan sisanya 0. Misalnya, jika ada 6 kelas, label "*age*" dapat diubah menjadi vektor [1, 0, 0, 0, 0, 0], "*ethnicity*" menjadi [0, 1, 0, 0, 0, 0], dan seterusnya. Pendekatan ini memungkinkan model untuk mengidentifikasi probabilitas setiap kelas dan sangat sesuai untuk tugas klasifikasi multikelas [2].

E. Pembagian Dataset

Setelah proses pra-pemrosesan dan tokenisasi, dataset dibagi menjadi tiga subset utama: set pelatihan (*training* set), set validasi (*validation set*), dan set pengujian (*test set*). Pembagian ini merupakan praktik standar dalam pengembangan model pembelajaran mesin untuk memastikan bahwa model dievaluasi secara objektif pada data yang belum pernah dilihat sebelumnya, sehingga memberikan indikasi yang akurat tentang kemampuan generalisasinya [2].

Berikut adalah fungsi dan peran masing-masing subset data:

- Set Pelatihan (*Training Set*): Merupakan bagian terbesar dari dataset yang digunakan untuk melatih model. Model mempelajari pola dan fitur dari data ini untuk melakukan tugas klasifikasi, menyesuaikan bobot internalnya untuk meminimalkan kesalahan prediksi.
- Set Validasi (*Validation Set*): Digunakan selama fase pelatihan untuk memantau kinerja model pada data yang belum terlihat. Set validasi ini krusial untuk melakukan penyesuaian *hyperparameter* (misalnya, *learning rate*, jumlah *epoch*) dan membantu dalam mendeteksi *overfitting*. Penggunaan set validasi

memungkinkan pemilihan model terbaik dari berbagai iterasi pelatihan tanpa harus "mencemari" *test set* [5].

- Set Pengujian (*Test Set*): Digunakan hanya sekali, setelah model final dipilih dan dilatih, untuk mengevaluasi kinerja akhir model secara independen dan tidak bias. Data dalam *test set* sama sekali tidak boleh digunakan selama pelatihan atau validasi untuk memastikan bahwa evaluasi kinerja model mencerminkan kemampuannya dalam menggeneralisasi pada data dunia nyata yang tidak terlihat sebelumnya [7].

Dalam penelitian ini, dataset dibagi dengan rasio sebagai berikut:

- 80% untuk set pelatihan
- 10% untuk set validasi
- 10% untuk set pengujian

Untuk memastikan representasi kelas yang proporsional di semua subset data, pembagian dataset dilakukan secara acak dengan teknik *stratified random sampling* berdasarkan label kelas (sebagaimana diimplementasikan menggunakan parameter stratify pada fungsi train_test_split dari pustaka scikit-learn). Dengan rasio 80% untuk set pelatihan, 10% untuk set validasi, dan 10% untuk set pengujian, strategi ini menjaga distribusi kelas asli yang relatif seimbang (sebagaimana ditunjukkan pada Tabel I) pada setiap subset, sehingga mengurangi potensi bias akibat ketidakseimbangan kelas dan mendukung evaluasi model yang lebih reliabel.

F. Arsitektur Model: RoBERTa-Large

Model inti yang digunakan dalam penelitian ini adalah RoBERTa-Large, sebuah arsitektur berbasis *Transformer encoder* yang telah dilatih sebelumnya (pre-trained) dan merupakan varian yang dioptimalkan dari *Bidirectional Encoder Representations from Transformers* (BERT) [12]. Model *Transformer* telah merevolusi bidang *Natural Language Processing* (NLP) dengan kemampuannya untuk memahami konteks dan dependensi jarak jauh dalam teks melalui mekanisme *self-attention* [12]. RoBERTa-Large secara khusus dirancang untuk meningkatkan kinerja BERT dengan modifikasi pada strategi pelatihan dan arsitektur, termasuk pelatihan pada korpus data yang lebih besar dan dengan *batch size* yang lebih besar, penghapusan operasi Next-Sentence Prediction (NSP), dan penggunaan teknik maskir dinamis [5].

Karakteristik utama RoBERTa-Large:

- Pre-trained Model**:** RoBERTa-Large adalah model yang telah dilatih pada dataset teks yang sangat besar (seperti CommonCrawl, Wikipedia, BookCorpus, dsb.) [5]. Proses *pre-training* ini memungkinkan model untuk mempelajari representasi bahasa yang kaya dan pengetahuan umum yang mendalam, sehingga model dapat memahami sintaksis, semantik, dan konteks kata secara efektif. Kemampuan ini menjadi fondasi yang kuat untuk tugas-tugas hilir (*downstream tasks*) seperti klasifikasi teks.
- Arsitektur Transformer Encoder**:** Model ini dibangun di atas arsitektur *encoder* dari *Transformer* asli. Ini terdiri dari beberapa lapisan *multi-head self-attention* dan jaringan *feed-forward*. Setiap lapisan memungkinkan model untuk mempertimbangkan semua token dalam urutan input secara bersamaan, menangkap hubungan kontekstual yang kompleks antar kata, tanpa batasan jarak seperti pada RNN [12].
- Bidirectional Context**:** Seperti BERT, RoBERTa dirancang untuk memahami konteks *bidirectional* (dua arah) dari teks. Artinya, setiap kata diwakili berdasarkan kata-kata di kiri dan kanannya, yang sangat penting untuk memahami makna penuh dari sebuah kalimat, terutama dalam konteks *cyberbullying* di mana nuansa dan konteks sangat penting [5].
- *Large Language Model* (LLM): RoBERTa-Large termasuk dalam kategori *Large Language Models* (LLMs) yang telah menunjukkan kinerja *state-of-the-art* pada berbagai tugas NLP [13]. Kemampuannya untuk menangani dan memproses informasi tekstual dalam skala besar membuatnya sangat cocok untuk tugas deteksi *cyberbullying* yang kompleks.

Dalam penelitian ini, RoBERTa-Large di-*fine-tune* untuk tugas klasifikasi *cyberbullying* multikelas. Ini berarti lapisan keluaran baru ditambahkan di atas model *pre-trained* RoBERTa-Large, dan seluruh model kemudian dilatih ulang (dengan *learning rate* yang rendah) pada dataset *cyberbullying* yang spesifik ini. Proses *fine-tuning* ini memungkinkan model untuk menyesuaikan pengetahuan umumnya dengan fitur-fitur spesifik dan nuansa bahasa dalam domain *cyberbullying*, sehingga menghasilkan kinerja klasifikasi yang optimal [5], [14].

G. Pendekatan Ensemble Learning

Untuk meningkatkan robusta dan kinerja deteksi *cyberbullying*, penelitian ini mengadopsi pendekatan *ensemble learning*. *Ensemble learning* adalah teknik di mana beberapa model pembelajaran mesin dilatih secara independen dan kemudian prediksi mereka digabungkan untuk menghasilkan prediksi akhir. Pendekatan ini seringkali menghasilkan kinerja yang lebih baik daripada model tunggal karena dapat mengurangi *bias* dan *variance*, serta meningkatkan stabilitas model secara keseluruhan [9]. Model *ensemble* cenderung lebih kuat terhadap *noise* dan

data yang tidak biasa karena mereka mengambil keuntungan dari kekuatan dan kelemahan masing-masing model konstituen [15], [16].

Dalam implementasi ini, tiga instans model RoBERTa-Large dilatih secara independen, dan strategi *soft voting* digunakan untuk menggabungkan prediksi mereka.

1. Pelatihan Model RoBERTa-Large Independen**:**

Tiga instans dari model RoBERTa-Large dilatih secara independen pada dataset pelatihan yang sama. Untuk memastikan variasi yang cukup antar model tanpa mengubah hyperparameter utama, setiap model dilatih dengan random seed yang berbeda: 42, 43, dan 44. Perbedaan dalam random seed ini memengaruhi inisialisasi bobot model dan urutan shuffling data pelatihan, yang pada gilirannya menghasilkan model-model yang sedikit berbeda dalam mempelajari pola dari data. Hal ini krusial untuk efektivitas pendekatan *ensemble*. Model RoBERTa-Large di-fine-tune dengan hyperparameter yang konsisten (misalnya, *learning rate*, *batch size*, *epoch*), kecuali untuk random seed yang berbeda untuk setiap instans. Untuk meningkatkan efisiensi komputasi selama proses pelatihan yang intensif ini, pelatihan setiap model dilakukan dengan memanfaatkan presisi campuran (*mixed precision*) FP16, yang memungkinkan penggunaan memori yang lebih rendah dan waktu pelatihan yang lebih cepat tanpa mengorbankan kinerja model secara signifikan.

2. Penggabungan Prediksi dengan *Soft Voting*

Setelah setiap dari tiga model RoBERTa-Large dilatih, mereka digunakan untuk membuat prediksi probabilitas untuk setiap kelas pada *test set*. Untuk setiap *tweet* dalam *test set*, setiap model menghasilkan serangkaian nilai probabilitas yang menunjukkan seberapa yakin model tersebut bahwa *tweet* tersebut termasuk dalam masing-masing dari enam kelas *cyberbullying* yang ada.

Strategi *soft voting* kemudian digunakan untuk menggabungkan hasil prediksi probabilitas ini. Caranya adalah dengan menghitung rata-rata probabilitas yang diprediksi oleh ketiga model untuk setiap kelas. Misalnya, jika Model 1 memprediksi probabilitas 0.8 untuk kelas "*age*", Model 2 0.7, dan Model 3 0.9, maka probabilitas *ensemble* untuk kelas "*age*" akan menjadi rata-rata dari ketiga nilai tersebut.

Prediksi akhir untuk setiap *tweet* adalah kelas yang memiliki nilai probabilitas rata-rata tertinggi dari semua kelas. Dengan kata lain, *ensemble* memilih kelas yang secara kolektif paling diyakini oleh ketiga model.

Pendekatan *soft voting* ini lebih disukai daripada *hard voting* (yang hanya mengambil kelas dengan mayoritas suara langsung dari setiap model) karena *soft voting* mempertimbangkan tingkat keyakinan (probabilitas) dari masing-masing model. Ini memberikan hasil yang lebih nuansa, mengakomodasi perbedaan tingkat keyakinan antar model, dan seringkali menghasilkan prediksi yang lebih stabil dan akurat [9].

Pemilihan strategi *soft voting* untuk menggabungkan beberapa model RoBERTa-Large yang dilatih secara independen dalam penelitian ini didasarkan pada kemampuannya untuk secara efektif memanfaatkan keragaman prediksi probabilistik yang dihasilkan oleh masing-masing model dasar yang sudah kuat. Mengingat bahwa setiap model RoBERTa-Large, meskipun memiliki arsitektur yang sama, dilatih dengan inisialisasi *seed* yang berbeda, masing-masing memiliki potensi untuk mempelajari nuansa dan pola data dari perspektif yang sedikit berbeda. Dengan merata-ratakan *output* probabilitas dari model-model ini, *soft voting* dapat menghasilkan keputusan *ensemble* yang lebih tergeneralisasi, mengurangi varians, dan meningkatkan stabilitas prediksi dibandingkan jika hanya mengandalkan satu model tunggal. Sementara metode *ensemble* lain seperti *boosting* (misalnya, AdaBoost, Gradient Boosting), yang secara iteratif melatih serangkaian model di mana setiap model baru difokuskan untuk memperbaiki kesalahan klasifikasi dari model sebelumnya, atau *stacking*, yang melibatkan pelatihan sebuah meta-model (atau *blender*) untuk mempelajari cara terbaik dalam menggabungkan prediksi dari berbagai model dasar (yang bisa jadi heterogen), juga dikenal sebagai pendekatan yang sangat efektif dalam banyak skenario.

Namun, kedua pendekatan tersebut seringkali menambah kompleksitas yang signifikan baik dalam implementasi maupun komputasi. *Boosting*, misalnya, bersifat sekuensial dan mungkin kurang intuitif untuk diterapkan secara langsung pada model *deep learning* yang sudah sangat dalam dan kompleks seperti RoBERTa-Large tanpa modifikasi yang cermat. Di sisi lain, *stacking* memerlukan pengelolaan dataset yang hati-hati untuk melatih meta-model (seringkali membutuhkan set data terpisah atau teknik validasi silang yang kompleks) untuk menghindari *overfitting*, dan kompleksitasnya meningkat seiring dengan jumlah dan keragaman model dasar. Oleh karena itu, *soft voting* dipilih dalam penelitian ini karena menawarkan keseimbangan yang sangat baik antara peningkatan kinerja prediktif, peningkatan robustisitas, dan kesederhanaan implementasi yang relatif tinggi, terutama ketika diterapkan pada *ensemble* model homogen (dalam hal ini, beberapa instans RoBERTa-Large) yang dilatih secara independen. Pendekatan ini juga secara inheren mendukung paralelisasi penuh dalam proses pelatihan model-model dasar, yang sangat efisien ketika berhadapan dengan model berukuran besar seperti RoBERTa-Large.

H. Konfigurasi Pelatihan dan Hiperparameter

Pelatihan setiap model RoBERTa-Large individual dilakukan dengan konfigurasi dan hiperparameter yang konsisten untuk memastikan perbandingan yang adil antar model sebelum di-*ensemble*. Pelatihan menggunakan

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 3, September 2025, Pp. 2666-2678

framework PyTorch dengan *TrainingArguments* dari pustaka Hugging Face Transformers. Hiperparameter utama yang digunakan meliputi: jumlah *epoch* pelatihan sebanyak 5, ukuran *batch* untuk pelatihan (*per_device_train_batch_size*) dan evaluasi (*per_device_eval_batch_size*) masing-masing 64, *warmup steps* sebanyak 200, dan *weight decay* sebesar 0.01. Untuk efisiensi komputasi dan mempercepat waktu pelatihan, digunakan teknik presisi campuran (*mixed precision*) dengan fp16=True. Proses evaluasi dilakukan pada setiap akhir *epoch* (eval_strategy="epoch"), dan model terbaik berdasarkan F1-*score* pada set validasi (metric_for_best_model="f1") disimpan secara otomatis (load_best_model_at_end=True). Setiap model individual dilatih dengan *seed* yang berbeda (42, 43, dan 44) untuk mendorong keragaman dalam *ensemble*.

III. HASIL DAN PEMBAHASAN

A. Hasil

Pelatihan tiga model RoBERTa-Large secara independen dengan *seed* yang berbeda, diikuti dengan penggabungan prediksi menggunakan strategi *soft voting*, telah dilakukan. Model ensemble yang dihasilkan kemudian dievaluasi secara komprehensif pada *test set* yang terdiri dari 4748 sampel *tweet*. Rincian distribusi sampel per kelas pada *test set* ini, termasuk nilai *support*nya, disajikan lebih lanjut pada Tabel III. Metrik evaluasi utama yang digunakan meliputi akurasi, presisi, recall, dan F1-score, yang merupakan metrik umum untuk menilai kinerja model klasifikasi [4].

1. Kinerja Keseluruhan Model Ensemble

Performa general dari model ensemble RoBERTa-Large dalam mengklasifikasikan *cyberbullying* multikelas pada *test set* disajikan pada Tabel II. Hasil ini mengindikasikan kapabilitas model dalam tugas yang diemban.

TABEL II
KINERJA KESELURUHAN MODEL ENSEMBLE (TEST SET)

Metrik	Nilai
Akurasi	0.8700
F1-Score (Weighted Avg)	0.8600
F1-Score (Macro Avg)	0.8600
Presisi (Weighted Avg)	0.8700
Recall (Weighted Avg)	0.8700

2. Analisis Kinerja Per Kelas

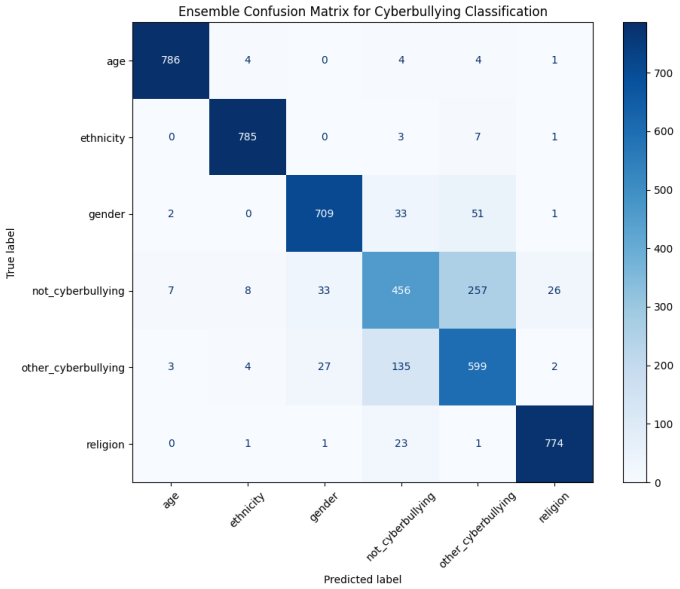
Untuk evaluasi yang lebih granular, kinerja model dianalisis untuk setiap kelas *cyberbullying*. Rincian nilai presisi, recall, dan F1-score untuk masing-masing dari enam kategori yang telah ditentukan dirangkum pada Tabel III.

TABEL III
KINERJA MODEL ENSEMBLE PER KELAS (TEST SET)

Kelas	Presisi	Recall	F1-Score	Support
Age	0.98	0.98	0.98	799
Ethnicity	0.98	0.99	0.98	796
Gender	0.92	0.89	0.91	796
Religion	0.96	0.97	0.96	800
Other_cyberbullying	0.65	0.78	0.71	770
Not_cyberbullying	0.70	0.58	0.63	787

Visualisasi dari distribusi prediksi model pada *test set* disajikan dalam bentuk *confusion matrix* pada Gambar 3 . *Confusion matrix* ini membantu dalam mengidentifikasi bagaimana model mengklasifikasikan setiap instance dan di mana misklasifikasi terjadi antar kelas. Terlihat bahwa model mencapai akurasi tinggi untuk kelas *age, ethnicity,* dan *religion*, yang tercermin dari konsentrasi nilai pada diagonal utama. Sebaliknya, misklasifikasi lebih sering terjadi pada kelas *not_cyberbullying* dan *other_cyberbullying*, yang mengindikasikan tantangan model dalam membedakan kedua kelas ini secara akurat.

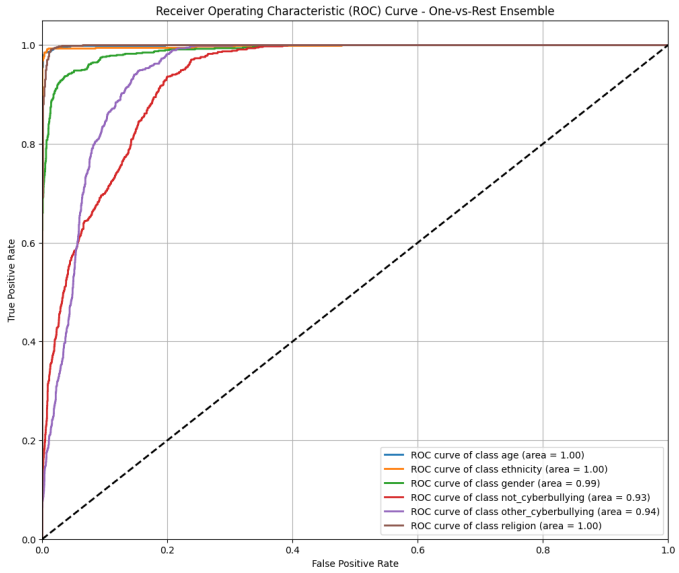
JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 3, September 2025, Pp. 2666-2678



Gambar. 3. Confusion matrix model ensemble pada test set

3. Analisis Kurva ROC AUC

Kemampuan model dalam membedakan antar kelas dievaluasi lebih lanjut menggunakan kurva *Receiver Operating Characteristic* (ROC) dan nilai *Area Under the Curve* (AUC), yang juga merupakan metrik penting dalam evaluasi model klasifikasi [4]. Gambar 4 menampilkan kurva ROC untuk setiap kelas dengan strategi *One-vs-Rest*.



Gambar. 4. Kurva ROC model ensemble dengan strategi One-vs-Rest

Nilai AUC yang dihasilkan untuk masing-masing kelas adalah sebagai berikut:

- age: 1.00
- ethnicity: 1.00
- gender: 0.99
- not_cyberbullying: 0.93
- other_cyberbullying: 0.94
- religion: 1.00 Nilai AUC yang mendekati 1.00 untuk sebagian besar kelas, khususnya *age, ethnicity, gender,* dan *religion*, menunjukkan kemampuan diskriminatif model yang sangat baik untuk kategori-kategori tersebut.

4. Metrik Proses Pelatihan Model

Proses pelatihan dari setiap model RoBERTa-Large individual dipantau untuk memastikan kualitas sebelum digabungkan ke dalam ensemble. Gambar 5 menyajikan kurva pembelajaran untuk model RoBERTa-Large

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 3, September 2025, Pp. 2666-2678

individual terakhir yang dilatih (menggunakan *seed* ke-3 atau seed 44) selama fase pelatihan.

Gambar. 5. Kurva pembelajaran model individual ke-3 (Seed 44)

Observasi pada Gambar 5 menunjukkan bahwa *training loss* untuk model ini secara konsisten menurun, sementara *validation loss* juga menunjukkan tren penurunan sebelum sedikit meningkat pada tahap akhir, yang mengindikasikan bahwa model telah belajar secara efektif dari data pelatihan dengan potensi *overfitting* yang relatif terkendali. Metrik F1-score dan akurasi pada set validasi juga meningkat secara stabil dan mencapai tingkat yang memuaskan.

Dua model RoBERTa-Large individual lainnya (yang dilatih dengan *seed* 42 dan 43) juga menunjukkan tren pelatihan dan validasi yang menghasilkan kinerja kompetitif. Tabel IV, yang merangkum metrik validasi utama setelah 5 epoch pelatihan, memberikan gambaran kuantitatif kinerja validasi akhir yang konsisten dari ketiga model individual tersebut.

TABEL IV
KINERJA VALIDASI AKHIR MODEL INDIVIDUAL (EPOCH 5)

Model (Seed)	Validation Loss (Akhir)	Validation F1-Score (Akhir)	Validation Accuracy (Akhir)
Model 1 (42)	0.410104	0.869689	0.870234
Model 2 (43)	0.370115	0.871484	0.871708
Model 3 (44)	0.398608	0.872457	0.872762

Data pada Tabel IV dan observasi pada Gambar 5 secara kolektif menunjukkan bahwa ketiga model individual telah dilatih dengan baik dan mencapai kinerja validasi yang kompetitif dan konsisten, sehingga layak untuk digabungkan dalam strategi ensemble. Pemantauan metrik pelatihan dan validasi seperti ini penting untuk menilai apakah model belajar dengan baik atau mengalami masalah seperti *overfitting* [4].

B. Pembahasan

Hasil eksperimen yang telah dipaparkan sebelumnya selanjutnya diinterpretasikan, dibandingkan dengan penelitian-penelitian terkait yang relevan, serta dianalisis dari sisi kekuatan, kelemahan, dan implikasinya.

1. Interpretasi Hasil Kinerja Model

Model ensemble RoBERTa-Large yang dikembangkan dalam penelitian ini berhasil mencapai akurasi keseluruhan 0.87 dan F1-Score (Weighted Avg) 0.86 pada *test set* (Tabel II). Kinerja ini menggarisbawahi efektivitas penggunaan model bahasa berbasis Transformer seperti RoBERTa, yang dikombinasikan dengan teknik *ensemble learning*, untuk tugas kompleks deteksi *cyberbullying* multikelas. Model *deep learning* memang seringkali menunjukkan keunggulan dalam menangani data tekstual bervolume besar dan melakukan ekstraksi fitur secara otomatis untuk tugas semacam ini [4].

Analisis kinerja per kelas (Tabel III) menunjukkan bahwa model sangat unggul dalam mengidentifikasi *cyberbullying* yang berkaitan dengan *age* (F1-score 0.98), *ethnicity* (F1-score 0.98), dan *religion* (F1-score 0.96). Nilai AUC yang sempurna (1.00) untuk ketiga kelas ini dan 0.99 untuk *gender* (Gambar 4) lebih lanjut menegaskan kemampuan diskriminatif model yang tinggi untuk kategori-kategori tersebut. Hal ini mengindikasikan bahwa representasi yang dipelajari oleh RoBERTa mampu menangkap ciri-ciri linguistik yang khas dan jelas untuk jenis-jenis *cyberbullying* ini.

Meskipun demikian, model menunjukkan kinerja yang lebih rendah untuk kelas *not_cyberbullying* (F1-score 0.63) dan *other_cyberbullying* (F1-score 0.71). Kesulitan dalam membedakan konten non-*cyberbullying* dari *cyberbullying*, atau mengklasifikasikan bentuk *cyberbullying* yang lebih umum dan tidak spesifik, merupakan

tantangan yang sering dilaporkan dalam literatur deteksi *cyberbullying* [4]. Secara lebih rinci, kelas *other_cyberbullying* kemungkinan besar berfungsi sebagai kategori '*catch-all*' yang menampung beragam bentuk perundungan siber yang tidak tergolong dalam kategori spesifik lainnya (seperti usia, etnis, gender, atau agama). Sifatnya yang sangat heterogen ini secara inheren menyulitkan model untuk mengidentifikasi pola linguistik yang konsisten dan unik, yang tercermin dalam nilai presisi (0.65) dan F1-score (0.71) yang lebih rendah (Tabel III). Observasi pada word cloud untuk kelas other_cyberbullying (Gambar 2) juga menunjukkan keragaman term yang luas, dengan beberapa kata kunci seperti '*bully*', '*people*', dan kata-kata umpatan ('*fucking*', '*idiot*') yang juga muncul dalam konteks *cyberbullying* spesifik lainnya atau bahkan dalam bahasa sehari-hari yang tidak merundung, sehingga memperkuat dugaan akan heterogenitas dan ambiguitas linguistik pada kelas ini. Sementara itu, untuk kelas *not_cyberbullying*, tantangan utama seringkali terletak pada kemiripan leksikal atau tematik dengan konten *cyberbullying* aktual, meskipun tanpa adanya niat merundung. Sebagai contoh, sebuah tweet mungkin berisi diskusi mengenai fenomena cyberbullying, mengutuk tindakan tersebut, atau mengajukan pertanyaan terkait, sehingga menggunakan kata-kata kunci yang serupa dengan yang ditemukan dalam ujaran cyberbullying itu sendiri. Hal ini juga tercermin pada *word cloud* kelas *not_cyberbullying* (Gambar 2), di mana kata '*bullying*' itu sendiri muncul secara dominan, bersama dengan term kontekstual seperti '*school*' dan '*people*'. Kehadiran term-term ini menunjukkan bahwa banyak sampel dalam kelas ini kemungkinan besar adalah diskusi atau laporan mengenai *cyberbullying*, bukan tindakan *cyberbullying* secara langsung. Model, meskipun secanggih RoBERTa-Large, mungkin masih menghadapi kesulitan dalam membedakan antara diskusi tentang *cyberbullying* dengan tindakan *cyberbullying* yang sesungguhnya, terutama jika hanya mengandalkan analisis teks per *tweet* tanpa konteks percakapan yang lebih luas atau pemahaman intensi pengguna yang mendalam. Hal ini berkontribusi pada nilai recall (0.58) dan F1-score (0.63) yang paling rendah di antara semua kelas (Tabel III).

Variasi bahasa yang luas, penggunaan sarkasme, ironi, dan konteks yang ambigu dalam kedua kategori ini, yaitu *not_cyberbullying* dan *other_cyberbullying*, kemungkinan menjadi faktor utama yang menyulitkan model. Confusion matrix (Gambar 3) juga mengilustrasikan adanya tumpang tindih prediksi yang signifikan antara kedua kelas ini, serta dengan beberapa kelas cyberbullying spesifik lainnya, yang menegaskan kompleksitas pembedaan ini. Untuk mengatasi tantangan pada kelas *not_cyberbullying* dan other_*cyberbullying* ini, beberapa pendekatan dapat dieksplorasi dalam penelitian selanjutnya. Misalnya, penerapan teknik augmentasi data yang ditargetkan untuk memperkaya variasi sampel pada kedua kelas sulit ini, atau penggunaan metode cost-sensitive learning yang memberikan bobot lebih pada kesalahan klasifikasi di kelas-kelas tersebut. Selain itu, eksplorasi fitur linguistik yang lebih canggih untuk menangkap sarkasme dan ambiguitas, serta potensi penyempurnaan definisi atau pemecahan kelas *other_cyberbullying* yang bersifat heterogen, dapat menjadi arah investigasi yang menjanjikan untuk peningkatan akurasi di masa depan.

2. Perbandingan dengan Penelitian Terkait

Hasil penelitian ini dapat dikontekstualisasikan dengan membandingkannya terhadap beberapa penelitian terkini yang relevan. Penggunaan model RoBERTa yang dimodifikasi dengan fitur tambahan seperti GloVe pada dataset *cyberbullying* publik yang serupa (Twitter, >47.000 tweet, 6 kelas) dilaporkan dapat mencapai akurasi 95% dan F1-score 96% [17]. Kinerja model dalam penelitian ini (akurasi 87%) yang lebih rendah mungkin mengindikasikan adanya potensi peningkatan melalui penambahan fitur semantik atau perbedaan dalam strategi *fine-tuning*.

Menariknya, pada dataset yang identik dengan yang digunakan dalam penelitian ini, pendekatan ensemble menggunakan model *machine learning* tradisional (Decision Trees, Random Forest, XGBoost) dengan fitur TF-IDF (bigram) mampu mencapai akurasi hingga 90.71% melalui *stacking classifier* [18]. Hal ini menunjukkan bahwa untuk dataset spesifik ini, pemilihan fitur dan arsitektur ensemble tertentu pada ML tradisional dapat sangat efektif, dan model *deep learning* yang lebih kompleks tidak secara otomatis menjamin keunggulan absolut tanpa optimasi yang cermat.

Strategi *ensemble learning* yang lebih kompleks seperti *stacking* juga telah menunjukkan hasil yang sangat menjanjikan dalam deteksi *cyberbullying* pada platform Twitter, bahkan mencapai akurasi 97.4% pada tugas klasifikasi biner dengan mengkombinasikan model seperti Conv1DLSTM, BiLSTM, LSTM, dan CNN [19]. Tingginya kinerja yang dilaporkan ini, meskipun pada skenario biner, mengisyaratkan adanya potensi untuk mengeksplorasi metode *ensemble* yang lebih beragam dibandingkan *soft voting* untuk penelitian ini guna peningkatan lebih lanjut.

Dalam konteks klasifikasi *cyberbullying* multikelas menggunakan data multi-modal (meme), model hybrid yang mengintegrasikan RoBERTa untuk analisis teks dengan model visi seperti ViT telah berhasil mencapai akurasi yang sangat tinggi, yaitu 99.20% pada dataset publik dan 96.10% pada dataset privat untuk 4 kelas *cyberbullying* [20]. Keberhasilan ini, terutama pada komponen teks yang ditangani oleh RoBERTa, mendukung relevansi

penggunaan RoBERTa untuk tugas multikelas dan sekaligus menyoroti potensi peningkatan kinerja jika modalitas data lain dapat diintegrasikan.

Dari perbandingan dengan penelitian-penelitian tersebut, terlihat bahwa lanskap deteksi *cyberbullying* sangat beragam, dengan berbagai pendekatan yang menunjukkan keunggulan pada skenario dan dataset tertentu. Beberapa penelitian [17], [19] melaporkan metrik kinerja yang lebih tinggi, seringkali dengan memanfaatkan fitur tambahan (seperti GloVe [17]), arsitektur *ensemble* yang lebih kompleks (seperti *stacking* [19]), atau fokus pada tugas klasifikasi biner yang mungkin memiliki tantangan berbeda dibandingkan klasifikasi multikelas. Penelitian lain [18] bahkan menunjukkan bahwa pada dataset yang identik dengan yang digunakan di sini, *ensemble* model *machine learning* tradisional dengan rekayasa fitur yang cermat dapat mencapai akurasi yang sangat kompetitif.

Meskipun demikian, signifikansi dari model *ensemble* RoBERTa-Large yang dikembangkan dalam penelitian ini terletak pada beberapa aspek. Pertama, penelitian ini mendemonstrasikan pencapaian kinerja yang solid dan kompetitif (Akurasi 0.87, F1-Score Weighted 0.86) untuk tugas deteksi *cyberbullying* multikelas yang kompleks dengan menggunakan pendekatan *end-to-end deep learning* berbasis *Transformer* murni, tanpa memerlukan rekayasa fitur manual ekstensif yang seringkali dibutuhkan oleh model *machine learning* tradisional. Kedua, konfigurasi *ensemble* dengan tiga model RoBERTa-Large yang dilatih dengan *seed* berbeda dan digabungkan melalui *soft voting*, terbukti mampu meningkatkan robustisitas dan stabilitas dibandingkan potensi model tunggal, sambil mempertahankan kompleksitas implementasi yang relatif lebih terkendali dibandingkan arsitektur *stacking* yang rumit. Ketiga, penelitian ini memberikan kontribusi pada pemahaman tentang bagaimana arsitektur RoBERTa-Large, bahkan tanpa fitur multimodal [20] atau modifikasi ekstensif, dapat menjadi fondasi yang kuat untuk deteksi *cyberbullying* multikelas dalam domain teks. Dengan demikian, model yang diusulkan menawarkan keseimbangan yang baik antara kinerja, kompleksitas implementasi, dan potensi generalisasi untuk aplikasi di dunia nyata, serta menyediakan *baseline* yang berharga untuk penelitian selanjutnya pada dataset dan tugas serupa.

3. Analisis Kekuatan dan Keterbatasan Model

Model yang dikembangkan dalam penelitian ini menunjukkan beberapa kekuatan signifikan. Salah satu yang paling menonjol adalah kemampuannya untuk mencapai kinerja sangat tinggi pada kelas-kelas *cyberbullying* spesifik seperti *age, ethnicity, religion,* dan *gender*, dengan nilai AUC yang mendekati 1.00 (Tabel III, Gambar 4), yang mengindikasikan kemampuan diskriminasi yang sangat baik untuk kategori-kategori tersebut. Kekuatan ini didukung oleh penggunaan RoBERTa-Large, sebuah model bahasa yang telah melalui *pre-training* pada korpus data yang sangat besar, sehingga menyediakan dasar representasi bahasa yang kaya akan konteks dan semantik, yang krusial untuk memahami nuansa dalam teks *cyberbullying* [4]. Lebih lanjut, pendekatan *ensemble learning* melalui *soft voting* dari tiga model individual dirancang untuk meningkatkan robustisitas dan stabilitas prediksi. Praktik ini umum digunakan untuk meningkatkan kinerja generalisasi model dengan mengurangi varians yang mungkin timbul dari satu proses pelatihan saja [19]. Dari sisi efisiensi, pelatihan dengan presisi campuran (FP16) juga menjadi keunggulan karena memungkinkan proses pelatihan model yang besar seperti RoBERTa-Large dapat dilakukan dengan kebutuhan sumber daya komputasi yang lebih optimal tanpa mengorbankan kinerja secara signifikan.

Meskipun demikian, model ini juga memiliki beberapa keterbatasan yang perlu menjadi perhatian. Teramati performa yang lebih rendah pada klasifikasi kelas *not_cyberbullying* dan *other_cyberbullying*. Hal ini menunjukkan bahwa model masih menghadapi kesulitan dalam menangani kasus-kasus yang lebih ambigu atau memiliki variasi linguistik yang lebih luas, sebuah tantangan yang memang sering dijumpai dalam upaya deteksi *cyberbullying* secara otomatis [4]. Selain itu, walaupun kinerja model secara keseluruhan tergolong baik, perbandingan dengan beberapa penelitian lain yang menggunakan dataset serupa menunjukkan bahwa masih terdapat ruang untuk peningkatan lebih lanjut. Beberapa penelitian lain berhasil mencapai akurasi yang sedikit lebih tinggi, kemungkinan melalui rekayasa fitur tambahan atau penggunaan strategi *ensemble* yang berbeda [17], [18]. Terakhir, seperti halnya banyak model *deep learning* lainnya, sifat "kotak hitam" dari arsitektur RoBERTa ensemble dapat membatasi tingkat interpretabilitas atau penjelasan atas keputusan yang dibuat oleh model, yang mana ini bisa menjadi pertimbangan penting dalam beberapa skenario aplikasi praktis [4].

IV. KESIMPULAN

Penelitian ini bertujuan untuk mengembangkan dan mengevaluasi sebuah sistem deteksi *cyberbullying* multikelas berkinerja tinggi dengan memanfaatkan kekuatan model bahasa besar RoBERTa-Large dalam konfigurasi ensemble dan dioptimalkan menggunakan presisi campuran. Metodologi yang diterapkan melibatkan *fine-tuning* tiga model RoBERTa-Large secara independen dengan *seed* yang berbeda pada dataset tweet berbahasa Inggris yang terdiri dari 47.692 sampel, yang dikategorikan ke dalam enam kelas *cyberbullying* yaitu *age, ethnicity,*

gender, religion, not_cyberbullying, dan *other_cyberbullying*. Prediksi dari ketiga model tersebut kemudian digabungkan menggunakan strategi *soft voting*.

Hasil evaluasi pada *test set* menunjukkan bahwa model ensemble yang diusulkan berhasil mencapai kinerja yang memuaskan dengan Akurasi sebesar 0.87 dan F1-Score (Weighted Avg) sebesar 0.86. Secara lebih rinci, model menunjukkan performa yang sangat baik dalam mengidentifikasi kelas-kelas *cyberbullying* spesifik seperti *age, ethnicity, gender,* dan *religion*, dengan nilai F1-score dan AUC yang tinggi untuk kategori-kategori tersebut. Meskipun demikian, model masih menghadapi tantangan dalam membedakan secara akurat kelas *not_cyberbullying* dan *other_cyberbullying*, yang menunjukkan adanya kompleksitas dan ambiguitas yang lebih tinggi pada kedua kelas tersebut. Proses pelatihan model individual juga menunjukkan kurva pembelajaran yang baik dengan potensi *overfitting* yang terkontrol, serta kinerja validasi yang konsisten antar model individual sebelum dilakukan ensemble.

Berdasarkan temuan-temuan tersebut, dapat disimpulkan bahwa pendekatan ensemble RoBERTa-Large dengan presisi campuran terbukti efektif untuk tugas deteksi *cyberbullying* multikelas berkinerja tinggi. Sistem yang dikembangkan tidak hanya mampu mengidentifikasi berbagai bentuk *cyberbullying* dengan akurasi yang baik tetapi juga menunjukkan potensi robustisitas melalui teknik *ensemble*. Penelitian ini memberikan kontribusi berupa sistem deteksi yang dapat menjadi dasar kuat untuk aplikasi praktis dalam upaya pencegahan dan penanganan *cyberbullying* di lingkungan digital. Untuk penelitian selanjutnya, disarankan agar fokus diberikan pada peningkatan kinerja untuk kelas-kelas yang masih menantang, eksplorasi fitur tambahan, pengujian strategi ensemble yang lebih beragam, serta pengembangan model untuk data multi-modal dan multibahasa guna meningkatkan generalisasi dan cakupan sistem.

DAFTAR PUSTAKA

[1] S. Bansal, N. Garg, J. Singh, and F. Van Der Walt, "Cyberbullying and mental health: past, present and future," *Front. Psychol.*, vol. 14, 2023, doi: 10.3389/fpsyg.2023.1279234.
[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] F. Elsafoury, S. Katsigiannis, Z. Pervez, and N. Ramzan, "When the Timeline Meets the Pipeline: A Survey on Automated Cyberbullying Detection," *IEEE Access*, vol. 9, pp. 103541–103563, 2021, doi: 10.1109/ACCESS.2021.3098979.
[4] 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," *Futur. Internet*, vol. 15, no. 5, pp. 1–47, 2023, doi: 10.3390/fi15050179.
[5] B. Ogunleye and B. Dharmaraj, "The Use of a Large Language Model for Cyberbullying Detection," *Analytics*, vol. 2, no. 3, pp. 694–707, 2023, doi: 10.3390/analytics2030038.
[6] H. Aljalaoud, K. Dashtipour, and A. Al-Dubai, "Arabic Cyberbullying Detection: A Comprehensive Review of Datasets and Methodologies," *IEEE Access*, vol. 13, no. March, pp. 69021–69038, 2025, doi: 10.1109/ACCESS.2025.3561132.
[7] Z. S. Bai and S. Malempati, "Ensemble Deep Learning (EDL) for Cyber-bullying on Social Media," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 7, pp. 551–560, 2023, doi: 10.14569/IJACSA.2023.0140761.
[8] Q. Li *et al.*, "A Survey on Text Classification: From Traditional to Deep Learning," *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 2, 2022, doi: 10.1145/3495162.
[9] S. Abimannan, E. S. M. El-Alfy, Y. S. Chang, S. Hussain, S. Shukla, and D. Satheesh, "Ensemble Multifeatured Deep Learning Models and Applications: A Survey," *IEEE Access*, vol. 11, no. September, pp. 107194–107217, 2023, doi: 10.1109/ACCESS.2023.3320042.
[10] A. Jakhotiya, H. Jain, B. Jain, and C. Chaniyara, "Text Pre-Processing Techniques in Natural Language Processing: A Review," *Int. Res. J. Eng. Technol.*, vol. 9, no. 2, pp. 878–880, 2022.
[11] S. Nazir, M. Asif, M. Rehman, and S. Ahmad, "Machine learning based framework for fine-grained word segmentation and enhanced text normalization for low resourced language," *PeerJ Comput. Sci.*, vol. 10, no. 1, pp. 1–19, 2024, doi: 10.7717/peerj-cs.1704.
[12] G. Tucudean, M. Bucos, B. Dragulescu, and C. D. Caleanu, "Natural language processing with transformers: a review," *PeerJ Comput. Sci.*, vol. 10, pp. 1–22, 2024, doi: 10.7717/PEERJ-CS.2222.
[13] Y. Chang *et al.*, "A Survey on Evaluation of Large Language Models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, 2024, doi: 10.1145/3641289.
[14] I. N. Santana, R. S. Oliveira, and E. G. S. Nascimento, "Text Classification of News Using Transformer-based Models for Portuguese," *J. Syst. Cybern. Informatics*, vol. 20, no. 5, pp. 33–59, 2022, doi: 10.54808/jsci.20.05.33.
[15] N. Alangari, M. El Bachir Menai, H. Mathkour, and I. Almosallam, "Exploring Evaluation Methods for Interpretable Machine Learning: A Survey," *Inf.*, vol. 14, no. 8, 2023, doi: 10.3390/info14080469.
[16] H. Allam, L. Makubvure, B. Gyamfi, K. N. Graham, and K. Akinwolere, "Text Classification: How Machine Learning Is Revolutionizing Text Categorization," *Inf.*, vol. 16, no. 2, 2025, doi: 10.3390/info16020130.
[17] A. A. Jamjoom, H. Karamti, M. Umer, S. Alsubai, T. H. Kim, and I. Ashraf, "RoBERTaNET: Enhanced RoBERTa Transformer Based Model for Cyberbullying Detection With GloVe Features," *IEEE Access*, vol. 12, no. May 2024, pp. 58950–58959, 2024, doi: 10.1109/ACCESS.2024.3386637.
[18] 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.
[19] A. Muneer, A. Alwadain, M. G. Ragab, and A. Alqushaibi, "Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT," *Inf.*, vol. 14, no. 8, 2023, doi: 10.3390/info14080467.
[20] I. Tabassum and V. Nunavath, "A Hybrid Deep Learning Approach for Multi-Class Cyberbullying Classification Using Multi-Modal Social Media Data," *Appl. Sci.*, vol. 14, no. 24, 2024, doi: 10.3390/app142412007.