KLASIFIKASI HIERARKIS UJARAN TOKSIK MULTI-LABEL MENGGUNAKAN MODEL RoBERTa

SKRIPSI

Oleh:
MUHAMMAD MUMTAZ SYAHRIE JAHIDY
NIM. 220605110072

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2026

KLASIFIKASI HIERARKIS UJARAN TOKSIK MULTI-LABEL MENGGUNAKAN MODEL RoBERTa

SKRIPSI

Diajukan kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)

Oleh:
MUHAMMAD MUMTAZ SYAHRIE JAHIDY
NIM. 220605110072

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2026

HALAMAN PERSETUJUAN

KLASIFIKASI HIERARKIS UJARAN TOKSIK MULTI-LABEL MENGGUNAKAN MODEL RoBERTa

SKRIPSI

Oleh:
MUHAMMAD MUMTAZ SYAHRIE JAHIDY
NIM. 220605110072

Telah Diperiksa dan Disetujui untuk Diuji:
Tanggal: 19 Juni 2026

Pembimbing I,	Pembimbing II,
Khadijah Fahmi Hayati Holle, M.Kom	Dr. Cahyo Crysdian, M.Cs
NIP. 19900626 202203 2 002	NIP. 19740424 200901 1 008

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Supriyono, M.Kom
NIP. 19841010 201903 1 012

HALAMAN PENGESAHAN

KLASIFIKASI HIERARKIS UJARAN TOKSIK MULTI-LABEL MENGGUNAKAN MODEL RoBERTa

SKRIPSI

Oleh:
MUHAMMAD MUMTAZ SYAHRIE JAHIDY
NIM. 220605110072

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 23 Juni 2026

Susunan Dewan Penguji

Ketua Penguji	: Dr. Zainal Abidin, M.Kom NIP. 19760613 200501 1 004	( )
Anggota Penguji I	: Tri Mukti Lestari, M.Kom NIP. 19911108 202012 2 005	( )
Anggota Penguji II	: Khadijah Fahmi Hayati Holle, M.Kom NIP. 19900626 202203 2 002	( )
Anggota Penguji III	: Dr. Cahyo Crysdian, M.Cs NIP. 19740424 200901 1 008	( )

Mengetahui dan Mengesahkan,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Supriyono, M.Kom
NIP. 19841010 201903 1 012

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama	: Muhammad Mumtaz Syahrie Jahidy
NIM	: 220605110072
Fakultas / Program Studi	: Sains dan Teknologi / Teknik Informatika
Judul Skripsi	: Klasifikasi Hierarkis Ujaran Toksik Multi-Label Menggunakan Model RoBERTa

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambil alihan data, tulisan, atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan skripsi ini merupakan hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 23 Juni 2026
Yang membuat pernyataan,

Muhammad Mumtaz Syahrie Jahidy
NIM. 220605110072

MOTTO

"Untuk mendapatkan yang belum pernah kamu dapatkan, kamu harus melakukan yang belum pernah kamu lakukan"

HALAMAN PERSEMBAHAN

Alhamdulillahi rabbil 'alamin, puji syukur senantiasa penulis panjatkan ke hadirat Allah *Subḥānahu wa Ta'ālā* atas segala limpahan rahmat, hidayah, serta karunia-Nya sehingga penyusunan skripsi ini dapat diselesaikan dengan baik. Shalawat seiring salam semoga terus tercurahkan kepada junjungan besar Nabi Muhammad *Ṣallallāhu 'Alaihi Wasallam* yang telah membimbing umat manusia menuju cahaya kebenaran dan peradaban addinul Islam yang mulia. Melalui lembar persembahan ini, penulis bermaksud menyampaikan rasa terima kasih dan penghormatan yang setinggi-tingginya kepada berbagai pihak yang telah menjadi pilar kekuatan selama menempuh pendidikan di jenjang perguruan tinggi.

Karya tulis ini secara khusus penulis persembahkan sebagai wujud bakti dan cinta kasih kepada kedua orang tua tercinta yang tidak pernah lelah merapalkan doa terbaik bagi kelancaran studi penulis. Dukungan moral, pengorbanan yang tak terhitung jumlahnya, serta kasih sayang mereka merupakan sumber energi utama yang membuat penulis sanggup melewati berbagai rintangan hingga mencapai tahap akhir ini. Selain itu, rasa terima kasih yang mendalam juga ditujukan kepada saudara-saudara penulis atas segala bentuk motivasi, semangat, serta bantuan yang terus mengalir tanpa henti sejak awal masa perkuliahan hingga draf skripsi ini benar-benar rampung.

Perjalanan akademik ini tentunya tidak akan terasa bermakna tanpa kehadiran orang-orang terdekat yang selalu mendampingi dalam setiap proses pembelajaran dan pendewasaan diri. Penulis mendedikasikan apresiasi ini kepada sahabat-sahabat masa Sekolah Menengah Atas yang terus menjaga tali silaturahmi,

serta teman-teman seperjuangan angkatan 2022 yang saling merangkul agar bisa lulus bersama. Tidak luput, ucapan terima kasih yang amat istimewa penulis sampaikan kepada keluarga besar kontrakan penghuni surga atas kebersamaan, canda tawa, dan kehangatan yang selalu menjadi tempat pulang paling nyaman di tengah penatnya menempuh revisi. Apresiasi yang sama juga penulis alamatkan kepada beberapa pihak lainnya yang tidak dapat disebutkan satu per satu, namun telah memberikan andil besar dalam kelancaran penyelesaian karya ilmiah ini.

Pada akhirnya, sebuah apresiasi yang paling jujur dan mendalam penulis tujukan untuk diri sendiri. Terima kasih karena telah memilih untuk terus bertahan, berjuang, dan menolak untuk menyerah di setiap fase yang penuh tekanan dan keraguan. Terima kasih atas segala kerja keras, keteguhan hati, serta malam-malam panjang yang telah dilalui di depan layar demi menyelesaikan tanggung jawab ini hingga tuntas ke garis akhir.

Penulis menyadari sepenuhnya bahwa keberhasilan menyelesaikan skripsi ini adalah hasil dari dukungan kolektif dan kebaikan hati banyak orang yang tidak akan mampu terbalas dengan sepadan. Oleh karena itu, penulis hanya dapat memanjatkan doa tulus agar seluruh bantuan, waktu, dan energi yang telah dicurahkan oleh semua pihak bernilai ibadah serta mendapatkan ganjaran kebaikan yang berlipat ganda dari Allah Subḥānahu wa Ta‘ālā. Semoga langkah kita semua ke depannya senantiasa diberkahi, dan karya sederhana ini diharapkan mampu memberikan manfaat yang nyata bagi perkembangan keilmuan, khususnya di bidang teknik informatika.

KATA PENGANTAR

Alhamdulillahi rabbil 'alamin, puji syukur senantiasa penulis panjatkan ke hadirat Allah *Subḥānahu wa Ta'ālā* atas segala limpahan rahmat, hidayah, serta kasih sayang-Nya yang telah memberikan kemudahan dan kelancaran bagi penulis untuk merampungkan karya ilmiah yang berjudul "KLASIFIKASI HIERARKIS UJARAN TOKSIK MULTI-LABEL MENGGUNAKAN MODEL RoBERTa". Shalawat seiring salam semoga terus tercurahkan kepada junjungan besar Nabi Muhammad *Ṣallallāhu 'Alaihi Wasallam*, teladan utama umat manusia yang syafaatnya selalu kita harapkan di hari akhir nanti, *Aamiin.*

Penulis menyadari sepenuhnya bahwa perjalanan panjang menyelesaikan skripsi ini tidak akan pernah terwujud tanpa adanya bimbingan, ketulusan, serta dukungan kolektif dari berbagai pihak. Oleh karena itu, dengan segala kerendahan hati, penulis ingin memberikan rasa terima kasih dan penghormatan yang setinggi-tingginya kepada:

1. Prof. Dr. Hj. Ilfi Nurdiana, M.Si., CAHRM., CRMP selaku Rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang, atas kesempatan dan fasilitas yang diberikan selama penulis menempuh pendidikan.
2. Dr. Agus Mulyono, M.Kes. selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang, atas segala dukungan dan kebijakan selama masa studi penulis.
3. Supriyono, M.Kom., selaku Ketua Program Studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.

4. Khadijah F.H. Holle, M.Kom., selaku Dosen Pembimbing I, yang telah melimpahkan kesabaran, waktu, serta bimbingan teoretis yang mendalam dari awal hingga akhir.
5. Dr. Cahyo Crysdian, M.Cs., selaku Dosen Pembimbing II, atas segala kebaikan, bimbingan teknis, dan saran akademis yang krusial dalam menyempurnakan skripsi ini.
6. Dr. Zainal Abidin, M.Kom., selaku Ketua Penguji, yang telah meluangkan waktu dan memberikan arahan serta evaluasi yang sangat berharga demi kesempurnaan skripsi ini.
7. Tri Mukti Lestari, M.Kom., selaku Anggota Penguji I, atas ketelitian, masukan ilmiah, dan kritik konstruktif yang sangat bermanfaat bagi penelitian ini..
8. Segenap Dosen, Staf Administrasi, dan Laboran, yang telah membekali penulis dengan ilmu pengetahuan serta memberikan bantuan layanan administratif selama masa perkuliahan
9. Kedua orang tua dan adik saya, pilar kekuatan utama yang tiada henti melangitkan doa, memberikan pengorbanan, dan motivasi agar penulis dapat menuntaskan studi dengan lancar.
10. Teman-teman sejawat dan sahabat seperjuangan angkatan 2022, rekan bertumbuh yang luar biasa, yang selalu menjadi teman diskusi yang kritis dan saling menguatkan dalam suka maupun duka.

Terakhir, sebuah apresiasi yang paling jujur penulis tujukan untuk diri sendiri. Terima kasih karena telah memilih untuk terus bertahan, berjuang, dan

menolak untuk menyerah di setiap fase yang melelahkan. Terima kasih atas segala kerja keras, keteguhan hati, serta malam-malam panjang yang telah dilalui di depan layar demi menyelesaikan tanggung jawab ini hingga ke garis akhir. Semoga karya tulis ini dapat membawa keberkahan dan manfaat bagi perkembangan keilmuan ke depannya.

Malang, 23 Juni 2026

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Jahidy, Muhammad Mumtaz Syahrie. 2026. **Klasifikasi Hierarkis Ujaran Toksik Multi-Label Menggunakan Model RoBERTa**. Skripsi. Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Khadijah Fahmi Hayati Holle, M.Kom. (II) Dr. Cahyo Crysdian, M.Cs.

Kata kunci: Ujaran Toksik, Klasifikasi Hierarkis, Klasifikasi Multilabel, RoBERTa, Ketidakseimbangan Kelas

Ujaran toksik di media sosial dapat memberikan dampak negatif terhadap individu maupun lingkungan digital sehingga diperlukan sistem deteksi otomatis yang mampu mengidentifikasi berbagai bentuk toksisitas secara akurat. Penelitian ini bertujuan menganalisis efektivitas pendekatan klasifikasi hierarkis berbasis RoBERTa dalam melakukan klasifikasi ujaran toksik multilabel pada teks berbahasa Indonesia. Penelitian membandingkan tiga skenario klasifikasi, yaitu klasifikasi biner independen, klasifikasi multilabel datar (flat multi-label classification), dan klasifikasi hierarkis (hierarchical classification). Untuk mengatasi permasalahan ketidakseimbangan kelas yang terdapat pada data, diterapkan teknik random undersampling pada proses pembagian data. Dataset dibagi menjadi data pelatihan dan data pengujian dengan rasio 85:15, sedangkan proses pelatihan dan validasi model dilakukan menggunakan Multilabel Stratified 5-Fold Cross Validation pada data pelatihan. Evaluasi dilakukan menggunakan metrik accuracy, precision, recall, F1-score, Subset Accuracy, dan hamming loss. Hasil pengujian menunjukkan bahwa model klasifikasi biner pada Skenario 1 memperoleh F1-score sebesar 79,07% dan recall sebesar 85,58%, sedangkan Level 1 pada pendekatan hierarkis memperoleh accuracy sebesar 79,64% dan F1-score sebesar 80,01%. Pada tugas klasifikasi multilabel, Skenario 2 menghasilkan macro *F1-score* sebesar 32,07%, sementara Level 2 pada Skenario 3 mencapai *macro F1-score* sebesar 52,47%. Evaluasi *end-to-end* menunjukkan bahwa pendekatan hierarkis menghasilkan *Subset Accuracy* sebesar 49,18%, lebih tinggi dibandingkan pendekatan *flat multi-label classification* yang memperoleh 43,68%. Hasil penelitian menunjukkan bahwa pendekatan klasifikasi hierarkis mampu meningkatkan performa klasifikasi ujaran toksik multilabel pada data yang tidak seimbang dengan membagi permasalahan ke dalam tahapan yang lebih terfokus. Temuan ini mengindikasikan bahwa struktur klasifikasi bertingkat lebih efektif dibandingkan pendekatan multilabel datar dalam menangani kompleksitas hubungan antarkategori ujaran toksik.

ABSTRACT

Jahidy, Muhammad Mumtaz Syahrie. 2026**. Hierarchical Multi-Label Toxic Speech Classification Using RoBERTa Model**. Undergraduate Thesis. Department of Informatics, Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang. Advisors: (I) Khadijah Fahmi Hayati Holle, M.Kom. (II) Dr. Cahyo Crysdian, M.Cs.

Keywords: Class Imbalance; Hierarchical Classification; Multi-Label Classification; RoBERTa; Toxic Speech

Toxic speech on social media can negatively affect individuals and digital communities, making automatic detection systems essential for identifying various forms of toxicity accurately. This study aims to analyze the effectiveness of a RoBERTa-based hierarchical classification approach for multi-label toxic speech classification in Indonesian texts. The research compares three classification scenarios, namely independent binary classification, flat multi-label classification, and hierarchical classification. To address the class imbalance problem within the dataset, a random undersampling technique was applied during split data process. The dataset was divided into training and testing sets using an 85:15 ratio, while model training and validation were conducted using Multilabel Stratified 5-Fold Cross Validation on the training set. Model performance was evaluated using accuracy, precision, recall, F1-score, Subset Accuracy, and hamming loss metrics. The experimental results show that the binary classification model in Scenario 1 achieved an F1-score of 79.07% and a recall of 85.58%, while Level 1 of the hierarchical approach obtained an accuracy of 79.64% and an F1-score of 80.01%. For the multi-label classification task, Scenario 2 achieved a macro F1-score of 32.07%, whereas Level 2 of Scenario 3 reached a macro F1-score of 52.47%. End-to-end evaluation demonstrated that the hierarchical approach achieved an Subset Accuracy of 49.18%, outperforming the flat multi-label classification approach, which achieved 43.68%. The findings indicate that hierarchical classification improves multi-label toxic speech classification performance on imbalanced data by decomposing the problem into more focused classification stages. These results suggest that a hierarchical classification structure is more effective than a flat multi-label approach in handling the complexity of relationships among toxic speech categories.

ملخص البحث

جاهيدي، محمد ممتاز شهري. 2026. **تصنيف الخطاب السام متعدد التصنيفات باستخدام النهج الهرمي ونموذج RoBERTa**.بحث جامعي. قسم المعلوماتية، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم الإسلامية الحكومية مالانج. المشرفان: (١) خديجة فهمي هياتي هولي، الماجستير. (٢) د. جاهيو كريسديان، الماجستير.

الكلمات المفتاحية: اختلال توازن الفئات؛ التصنيف الهرمي؛ التصنيف متعدد التصنيفات؛RoBERTa ؛ الخطاب السام

يمكن أن يؤثر الخطاب السام على وسائل التواصل الاجتماعي تأثيرًا سلبيًا على الأفراد والمجتمعات الرقمية، مما يجعل أنظمة الكشف التلقائي ضرورةً أساسيةً للتعرف على أشكال السمية المختلفة بدقة عالية. تهدف هذه الدراسة إلى تحليل فاعلية نهج التصنيف الهرمي المبني على نموذج RoBERTa لتصنيف الخطاب السام متعدد التصنيفات في النصوص الإندونيسية. تقارن الدراسة ثلاثة سيناريوهات للتصنيف، هي: التصنيف الثنائي المستقل، والتصنيف المسطّح متعدد التصنيفات، والتصنيف الهرمي. ولمعالجة مشكلة اختلال توازن الفئات في مجموعة البيانات، تم تطبيق أسلوب التقليل العشوائي للعينات خلال عملية التدريب. قُسِّمت مجموعة البيانات إلى مجموعتَي تدريب واختبار بنسبة 85:15، فيما أُجري تدريب النموذج والتحقق من صحته باستخدام التحقق المتقاطع متعدد التصنيفات الطبقي بخمسة أجزاء على مجموعة التدريب. وقد قُيِّم أداء النموذج باستخدام مقاييس الدقة، والضبط، والاسترجاع، ومقياسF1 ، ودقة المجموعة الفرعية، وخسارة هامينج. أظهرت نتائج التجارب أن نموذج التصنيف الثنائي في السيناريو الأول حقق مقياس F1 بنسبة 79.07% ومعدل استرجاع 85.58%، في حين حقق المستوى الأول من النهج الهرمي دقةً بلغت 79.64% ومقياس F1 بنسبة 80.01%. أما على صعيد مهمة التصنيف متعدد التصنيفات، فقد حقق السيناريو الثاني مقياس F1 الكلي بنسبة 32.07%، بينما بلغ المستوى الثاني من السيناريو الثالث مقياس F1 الكلي 52.47%. وكشف التقييم الشامل من البداية إلى النهاية أن النهج الهرمي حقق دقة مجموعة فرعية بلغت 49.18%، متفوقًا بذلك على نهج التصنيف المسطّح متعدد التصنيفات الذي حقق 43.68%. تشير هذه النتائج إلى أن التصنيف الهرمي يُحسِّن أداء تصنيف الخطاب السام متعدد التصنيفات على البيانات غير المتوازنة، وذلك بتحليل المشكلة إلى مراحل تصنيف أكثر تركيزًا. وتدل هذه النتائج على أن البنية الهرمية للتصنيف أكثر فاعليةً من النهج المسطّح متعدد التصنيفات في التعامل مع تعقيد العلاقات بين فئات الخطاب السام.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Media sosial telah menjadi salah satu sarana utama bagi masyarakat Indonesia untuk berinteraksi, berdiskusi, dan menyampaikan pendapat. Platform seperti *Twitter* (sekarang *X*), Instagram, dan TikTok tidak hanya berfungsi sebagai tempat hiburan, tetapi juga sebagai kanal informasi dan komunikasi yang sangat penting dalam kehidupan sosial-politik (Santoso dkk., 2020). Dalam konteks ideal, media sosial seharusnya menjadi ruang bagi demokrasi digital, memperkuat keterbukaan, dan memperluas dialog publik yang sehat (Cohen & Fung, 2023).

Namun, realitas di lapangan menunjukkan adanya permasalahan serius dalam komunikasi digital di Indonesia, yaitu meningkatnya ujaran toksik di media sosial. Ujaran toksik mengacu pada bentuk komunikasi yang berisi penghinaan, kebencian, provokasi, atau ekspresi negatif yang bertujuan untuk merendahkan, menyerang, atau mencemarkan pihak lain (Amin dkk., 2018). Fenomena ini terutama terlihat dalam diskursus politik yang seringkali bersifat polarisasi, menciptakan iklim komunikasi yang tidak sehat dan memicu konflik sosial (Sukidin dkk., 2025). Ujaran toksik tersebut tidak hanya berdampak pada individu yang menjadi sasaran, tetapi juga pada masyarakat secara keseluruhan, memperburuk kualitas diskusi publik dan merusak ruang komunikasi yang konstruktif.

Salah satu tantangan terbesar dalam mendeteksi ujaran toksik di media sosial adalah keberagaman cara penyampaian pesan toksik, yang seringkali tidak

secara eksplisit menyertakan kata-kata kasar atau penghinaan. Banyak ujaran toksik yang disampaikan secara tersirat, seperti melalui sindiran, makna ganda, atau plesetan, yang seringkali bergantung pada konteks sosial dan budaya tertentu (Das & Balke, 2024). Bentuk-bentuk ujaran seperti ini lebih sulit untuk dideteksi menggunakan metode tradisional, seperti pencocokan kata kunci, karena mereka tidak selalu mengandung kata negatif secara eksplisit, meskipun tetap dapat merendahkan atau menghina pihak lain (Zhou dkk., 2025). Ini menunjukkan bahwa teknologi yang ada saat ini belum memadai dalam memahami konteks dan kompleksitas bahasa yang digunakan di media sosial.

Dalam konteks ini, penelitian tentang deteksi ujaran toksik di media sosial, khususnya dalam bahasa Indonesia, menjadi sangat penting. Terlebih lagi, data yang digunakan untuk penelitian ini seringkali sangat tidak seimbang, di mana jumlah data non-toksik jauh lebih banyak dibandingkan dengan data toksik. Hal ini menciptakan tantangan tersendiri dalam melatih model yang efektif untuk mendeteksi ujaran toksik dengan akurasi tinggi. Oleh karena itu, diperlukan pendekatan yang lebih canggih untuk memecahkan masalah ini, baik dari segi *struktur model* maupun *pemahaman konteks*.

Penelitian terdahulu (Susanto, Wijanarko, Pratama, Hong, dkk., 2025) yang menggunakan model berbasis *BERT* atau *IndoBERTweet* untuk mendeteksi ujaran kebencian cenderung terbatas pada klasifikasi biner. Namun, dalam banyak kasus, satu teks dapat mengandung lebih dari satu jenis ujaran kebencian, seperti ujaran kebencian yang juga mengandung ancaman atau penghinaan. Oleh karena itu, diperlukan pendekatan yang lebih kompleks, seperti *klasifikasi multi-label*, yang

tidak hanya mendeteksi keberadaan ujaran toksik, tetapi juga mengidentifikasi jenis toksisitas yang terkandung di dalamnya. Selain itu, *pendekatan hierarkis* dapat membantu mengatasi masalah ketidakseimbangan data, di mana model dapat lebih dulu memfilter data non-toksik dan kemudian fokus pada klasifikasi jenis toksisitas dari data yang lebih sedikit.

Model *RoBERTa* dipilih dalam penelitian ini karena kemampuannya dalam menangani konteks yang lebih dalam dan kompleks dibandingkan dengan model-model sebelumnya, seperti *BERT*. *RoBERTa* memiliki keunggulan dalam hal *pre-training*, dengan menggunakan teknik *dynamic masking* dan *korpus pelatihan yang lebih besar*, yang memungkinkan model ini untuk lebih adaptif terhadap variasi sintaksis dan semantik dalam bahasa alami. Dalam konteks deteksi ujaran toksik, kemampuan *RoBERTa* untuk memahami konteks yang lebih dalam sangat penting untuk membedakan antara teks yang hanya berisi kata-kata biasa dengan teks yang mengandung makna toksik yang lebih tersirat. Dalam ajaran Islam, pelaku ujaran toksik dikutuk oleh Allah sebagaimana Q.S. Al-Humazah: 1.

وَيْلٌ لِّكُلِّ هُمَزَةٍ لُّمَزَةٍ ﴿١﴾

"Celakalah bagi setiap pengumpat dan pencela." (Q.S. Al-Humazah: 1)

Tafsir singkat dari ayat tersebut menjelaskan bahwa Allah Swt. memberikan peringatan kepada setiap orang yang gemar menghina, mencela, atau menyakiti orang lain melalui ucapan maupun perkataan buruk. Perilaku tersebut dapat menimbulkan permusuhan, kebencian, dan kerusakan dalam hubungan sosial antarmanusia.

Berdasarkan kondisi tersebut, maka diusulkan penelitian dengan judul Klasifikasi Hierarkis Ujaran Toksik Multi-Label Menggunakan Model RoBERTa melalui pengukuran metrik performa seperti *precision*, *recall*, *F1-score*, *accuracy*, dan *Hamming loss*, guna menilai kualitas klasifikasi ujaran toksik dalam diskursus politik Indonesia di media sosial. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi bagi pengembangan sistem analisis ujaran toksik berbahasa Indonesia yang lebih presisi dan kontekstual, sekaligus mendukung terciptanya ruang komunikasi digital yang lebih sehat dan konstruktif di ranah digital Indonesia.

1.2 Pernyataan Masalah

Berdasarkan uraian latar belakang di atas, maka permasalahan yang menjadi fokus penelitian ini adalah bagaimana performa model *FacebookAI/xlm-roberta-base* pada klasifikasi ujaran toksik berbahasa Indonesia dalam tiga skenario, yaitu klasifikasi biner, flat multi-label classification, dan hierarchical multi-label classification yang diukur menggunakan metrik performa seperti *precision*, *recall*, *F1-score***,** *Macro-F1, accuracy*, *Subset Accuracy* dan *Hamming loss*?

1.3 Batasan Masalah

Untuk menjaga fokus penelitian dan menghindari perluasan ruang lingkup yang berlebihan, penelitian ini dibatasi pada beberapa hal sebagai berikut:

1. Analisis terbatas pada ujaran toksik yang muncul di media sosial, khususnya pada platform Twitter (X). Data yang digunakan berasal dari dataset

IndoDiscourse yang telah terkurasi dan mewakili berbagai jenis ujaran toksik berbahasa Indonesia.

2. Pendekatan klasifikasi yang digunakan ada 3, klasifikasi biner, klasifikasi *flat multilabel*, dan klasifikasi hierarkis. Untuk klasifikasi hierarkis terdiri dari dua level analisis, yaitu:
   a) Level 1: Deteksi teks toksik atau non-toksik (klasifikasi biner).
   b) Level 2: *Klasifikasi multi-label* terhadap teks yang teridentifikasi sebagai *toxic*, untuk menentukan jenis toksisitas meliputi *polarized, profanity_obscenity, threat_incitement_to_violence, insults, identity_attack, sexually_explicit*.
3. Evaluasi performa model dibatasi pada pengukuran metrik *precision*, *recall*, *F1-score*, *accuracy*, *Subset Accuracy, E2E Accuracy* dan *Hamming loss* untuk menilai performa model klasifikasi.
4. Model yang dipakai adalah *FacebookAI/xlm-roberta-base*.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk mengevaluasi performa pendekatan klasifikasi ujaran toksik multi-label menggunakan model *FacebookAI/xlm-roberta-base* dalam tiga skenario, yaitu klasifikasi biner, flat multi-label classification, dan hierarchical multi-label classification melalui pengukuran metrik performa seperti *precision*, *recall*, *F1-score*, *Macro-F1, accuracy*, *Subset Accuracy* dan *Hamming loss*, guna menilai kualitas klasifikasi ujaran toksik di media sosial.

1.5 Manfaat Penelitian

Penelitian ini memiliki manfaat baik secara teoretis maupun praktis. Secara teoretis, penelitian ini diharapkan dapat memperkaya pengembangan ilmu di bidang *Natural Language Processing (NLP)*, khususnya terkait penerapan model *deep learning* untuk klasifikasi ujaran toksik dalam bahasa Indonesia. Pendekatan klasifikasi hierarkis ujaran toksik multi-label yang diusulkan dapat menjadi kontribusi baru dalam memperluas pemahaman terhadap struktur dan hubungan antarjenis toksisitas dalam teks berbahasa alami. Hasil penelitian ini juga diharapkan menjadi rujukan bagi penelitian-penelitian selanjutnya yang berfokus pada klasifikasi ujaran toksik, analisis sentimen, atau klasifikasi teks kontekstual dalam bahasa Indonesia.

Dari sisi praktis, hasil penelitian ini dapat dimanfaatkan oleh berbagai pihak, seperti lembaga pemerintah, media massa, maupun platform media sosial, untuk mengembangkan sistem moderasi konten otomatis yang lebih kontekstual dan sensitif terhadap budaya Indonesia. Dengan kemampuan menklasifikasi ujaran toksik secara lebih akurat, sistem ini dapat membantu mengurangi penyebaran ujaran toksik dan memperbaiki kualitas diskusi politik di ruang digital.

Secara sosial, penelitian ini diharapkan dapat berkontribusi dalam menciptakan lingkungan komunikasi publik yang lebih sehat, etis, dan toleran. Klasifikasi otomatis terhadap ujaran toksik tidak hanya penting untuk menjaga etika berkomunikasi di dunia maya, tetapi juga sebagai langkah preventif terhadap potensi konflik sosial yang dipicu oleh ujaran toksik.

BAB II

STUDI PUSTAKA

2.1 Penelitian Terkait

Penelitian mengenai ujaran toksik di media sosial terus berkembang seiring meningkatnya interaksi publik dalam ruang digital, khususnya pada isu-isu politik di Indonesia. Sejumlah studi telah dilakukan dengan memanfaatkan teknik *Natural Language Processing* (NLP) dan *Deep Learning* untuk mengenali ujaran yang bersifat ofensif atau diskriminatif dalam teks berbahasa Indonesia.

Handayani dan Gani (2023) mengusulkan model *Recurrent Neural Network (LSTM)* dengan penalaan hiperparameter untuk mendeteksi ujaran kebencian di Twitter dan memperoleh akurasi 92.4% serta F1-score 91.8%. (Susanto, Wijanarko, Pratama, Hong, dkk., 2025) melalui *IndoDiscourse* memperkenalkan dataset baru dengan 43.692 entri beranotasi demografis dan menghasilkan F1-score 0.78 menggunakan *IndoBERTweet*. (Findawati dkk., 2023) menerapkan *ensemble classifier* berbasis *IndoBERTweet* untuk klasifikasi multi-label pada data tidak seimbang dengan akurasi 90.2%. Sementara itu, (Azhari dkk., 2023) menggunakan model *RoBERTa* untuk mendeteksi ujaran kebencian pada komentar Instagram dan mencapai akurasi 85% dengan F1-score 84.9%.

Tabel 2.1 merangkum empat penelitian terkait yang relevan dengan topik ini, dengan fokus pada metodologi, hasil utama, serta kontribusi terhadap pengembangan sistem klasifikasi ujaran toksik berbahasa Indonesia.

Tabel 2.1 Penelitian Terkait

No	Peneliti	Judul	Metode	Hasil
1	(Van Aken dkk., 2018)	*Challenges for Toxic Comment Classification: An In-Depth Error Analysis*	Analisis mendalam klasifikasi toxic comment dengan beberapa pendekatan dan error analysis	deteksi toxic comment masih menghadapi tantangan besar, terutama missing context dan inconsistent labels
2	(Ibrohim & Budi, 2019)	*Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter*	SVM, NB, RFDT serta BR, LP, CC	hate speech Indonesia lebih tepat dipandang sebagai multi-label problem karena ada target, category, dan level
3	(Koto dkk., 2021)	*INDOBERTWEET: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization*	Pengembangan IndoBERTweet	data Twitter Indonesia memerlukan model yang peka terhadap domain-specific vocabulary dan mismatch kosakata
4	(L. Xu dkk., 2021)	*Hierarchical Multi-label Text Classification with Horizontal and Vertical Category Correlations*	Hierarchical Multi-label Text Classification dengan korelasi vertikal dan horizontal antar kategori	banyak studi mereduksi masalah hierarkis menjadi flat multi-label, sehingga terjadi information loss
5	(Handayani & Hilmansyah Gani, 2023)	*Enhancing Multi-Label Hate Speech and Abusive Language Detection on Indonesian Twitter Using Recurrent Neural Networks with Hyperparameter Tuning*	RNN (LSTM)	Akurasi 92.4%, F1-score 91.8% setelah tuning embedding = 32, LSTM units = 32, dropout = 0.2
6	(Findawati dkk., 2023)	*Multi-Label Aspect Dangerous Speech Classification Using Keyword-Driven Ensemble Classifier on Imbalanced Data*	*Ensemble classifier* berbasis keyword + *IndoBERTweet*	Akurasi 90.2%, F1-score 89.5% pada data imbalanced
7	(Azhari dkk., 2023)	*Detection of Indonesian Hate Speech in the Comments Column of Indonesian Artists' Instagram Using the RoBERTa Method*	*RoBERTa*	Akurasi 85%, precision 84.6%, recall 85.2%, F1-score 84.9%
8	(Susanto, Wijanarko, Pratama, Tang, dkk., 2025)	*A Multi-Labeled Dataset for Indonesian Discourse: Examining Toxicity, Polarization, and Demographics Information*	*model BERT-base* dan *LLMs*	Model IndoBERTweet-base mencapai F1 = 0,812 pada deteksi toksisitas dan F1 = 0,794 pada polarization; GPT-3.5 mencapai F1 = 0,833 dan 0,817

Penelitian pada Tabel 2.1 menunjukkan penelitian mengenai ujaran toksik telah berkembang dari pendekatan klasifikasi biner menuju pemodelan yang lebih kompleks karena satu ujaran sering kali tidak hanya mengandung satu bentuk pelanggaran, tetapi dapat memuat beberapa dimensi toksisitas secara bersamaan. (Van Aken dkk., 2018) menunjukkan bahwa klasifikasi ujaran toksik masih menghadapi tantangan penting, seperti keterbatasan konteks dan inkonsistensi label, sehingga pendekatan yang terlalu sederhana belum memadai untuk merepresentasikan kompleksitas ujaran toksik di media sosial.

Dalam konteks bahasa Indonesia, (Ibrohim & Budi, 2019) menunjukkan bahwa *hate speech* dan *abusive language* lebih tepat dipandang sebagai masalah multi-label, karena satu teks dapat memiliki lebih dari satu atribut sekaligus, seperti target, kategori, dan level ujaran. Temuan ini menjadi fondasi penting bahwa klasifikasi ujaran bermasalah di media sosial Indonesia tidak cukup diselesaikan dengan skema label tunggal. Namun, penelitian tersebut masih didominasi pendekatan machine learning klasik dan belum memanfaatkan representasi kontekstual berbasis transformer secara mendalam.

Perkembangan berikutnya ditandai oleh hadirnya model bahasa yang dirancang khusus untuk media sosial Indonesia. (Koto Jey Han Lau Timothy Baldwin, 2021) memperkenalkan IndoBERTweet sebagai model pralatih untuk Twitter Indonesia dan menunjukkan pentingnya adaptasi kosakata domain-spesifik agar model mampu menangani karakteristik bahasa informal, singkatan, variasi ejaan, dan gaya komunikasi khas media sosial. Temuan ini memperkuat bahwa

penelitian ujaran toksik di platform seperti Twitter/X membutuhkan model kontekstual yang sensitif terhadap domain bahasa Indonesia digital.

Selain persoalan konteks bahasa, tantangan lain yang dominan adalah ketidakseimbangan distribusi label dan tumpang tindih antarlabel. Keempat penelitian terakhir menunjukkan perkembangan bertahap dalam upaya memahami dan mengklasifikasikan ujaran toksik di media sosial berbahasa Indonesia. Penelitian oleh (Findawati dkk., 2023) menjadi salah satu landasan penting dalam penanganan data berlabel ganda dengan menerapkan *ensemble classifier* untuk mengatasi ketidakseimbangan data. Pendekatan ini membuka arah bagi penelitian berikutnya untuk memanfaatkan model berbasis *deep learning* yang lebih kontekstual.

Kusuma dan Chowanda (2023) memperkuat pendekatan tersebut melalui kombinasi *IndoBERTweet* dan *BiLSTM* yang mampu memahami konteks kalimat secara mendalam dan menghasilkan akurasi yang tinggi dalam mendeteksi ujaran kebencian di Twitter. Selanjutnya, (Azhari dkk., 2023) menggunakan model *RoBERTa* untuk mendeteksi ujaran kebencian pada komentar Instagram. Hasilnya menunjukkan bahwa model berbasis *transformer* mampu menangkap makna semantik yang lebih kompleks dibandingkan pendekatan sebelumnya.

Penelitian *IndoDiscourse* (Susanto, Wijanarko, Pratama, Tang, dkk., 2025) memperluas lingkup analisis dengan memperkenalkan dataset beranotasi multi-label yang mencakup dimensi toksisitas dan polarisasi dalam konteks politik Indonesia. Studi ini menunjukkan peningkatan performa model dengan *macro F1-score* mencapai 0,83 menggunakan *IndoBERTweet* dan *GPT-4o-mini*. Namun,

penelitian tersebut belum menerapkan pendekatan klasifikasi hierarkis yang membedakan tahap deteksi toksisitas dan tahap klasifikasi jenis toksisitas secara berurutan.

Berdasarkan penelitian-penelitian tersebut, masih terdapat kesenjangan dalam pemodelan ujaran toksik berbahasa Indonesia, khususnya dalam hal struktur label yang bersifat hierarkis. Pendekatan yang ada umumnya hanya mendeteksi ujaran toksik secara umum tanpa mengklasifikasikan lebih lanjut ke dalam kategori toksisitas seperti *polarized*, *profanity_obscenity*, *threat_incitement_to_violence*, *insults*, *identity_attack*, dan *sexually_explicit*. Selain itu, belum ada penelitian yang mengintegrasikan pendekatan hierarkis menggunakan model *transformer* modern seperti *RoBERTa* yang memiliki kemampuan kontekstual lebih baik dalam memahami relasi antar-label dalam satu wacana.

Oleh karena itu, penelitian ini diusulkan untuk mengembangkan dan mengevaluasi klasifikasi hierarkis ujaran toksik multi-label menggunakan model. Penelitian ini bertujuan untuk menilai performa model berdasarkan metrik evaluasi *accuracy*, *precision*, *recall*, *F1-score*, dan *Hamming loss*, sehingga diharapkan dapat menghasilkan sistem klasifikasi yang lebih akurat dan adaptif terhadap kompleksitas bahasa media sosial Indonesia.

2.2 Research Gap

Berdasarkan perkembangan tersebut, terdapat beberapa celah penelitian yang masih terbuka. Pertama, penelitian ujaran toksik berbahasa Indonesia memang telah bergerak dari klasifikasi biner menuju multi-label, tetapi sebagian besar masih menggunakan pendekatan *flat classification* atau model klasik, sehingga belum

sepenuhnya memanfaatkan struktur label yang bertingkat. Kedua, penelitian yang menangani data multi-label tidak seimbang sudah mulai berkembang, tetapi umumnya masih berfokus pada kombinasi fitur atau *ensemble classifier*, bukan pada skema klasifikasi hierarkis dua tahap berbasis transformer kontekstual.

Ketiga, meskipun IndoDiscourse telah menyediakan dataset yang mutakhir dan sangat relevan untuk konteks Indonesia, benchmark yang tersedia belum secara khusus menguji pendekatan *two-stage hierarchical classification* yang memisahkan tahap deteksi *toxic/non-toxic* dari tahap klasifikasi jenis toksisitas. Padahal, secara konseptual, identifikasi jenis toksisitas lebih logis dilakukan setelah suatu teks terlebih dahulu dinyatakan toksik. Keempat, literatur HMTC menunjukkan bahwa pengabaian struktur hierarkis berpotensi menyebabkan kehilangan informasi dependensi antarlabel, sehingga masih diperlukan penelitian yang secara eksplisit menggabungkan struktur label bertingkat, *multi-label predict*ion, dan model transformer dalam satu kerangka eksperimen.

Dengan demikian, gap utama penelitian ini terletak pada belum adanya kajian yang secara spesifik menerapkan *hierarchical multi-label classification* menggunakan RoBERTa pada data ujaran toksik bahasa Indonesia yang berstruktur bertingkat, sekaligus mengevaluasi efektivitas pendekatan tersebut dalam menghadapi konteks bahasa media sosial, ketidakseimbangan label, dan kemunculan lebih dari satu jenis toksisitas dalam satu teks

2.3 Ujaran Toksik dan Ujaran Kebencian di Media Sosial

Ujaran toksik merupakan bentuk komunikasi daring yang mengandung unsur penghinaan, provokasi, diskriminasi, atau serangan terhadap individu

maupun kelompok tertentu (Bate, 2024). Ujaran toksik umumnya ditandai oleh bahasa yang kasar, merendahkan, atau menyudutkan pihak lain tanpa harus selalu menargetkan identitas sosial tertentu. Sementara itu, ujaran kebencian secara lebih spesifik mengandung ekspresi kebencian yang ditujukan kepada suatu kelompok berdasarkan ras, agama, jenis kelamin, etnisitas, orientasi politik, atau identitas sosial lainnya. Perbedaan ini dijelaskan oleh (Fortuna & Nunes, 2018) bahwa ujaran kebencian bersifat ideologis dan kolektif karena menyerang identitas kelompok, sedangkan ujaran toksik lebih bersifat interpersonal, menyerang individu atau opini tertentu tanpa selalu mengandung unsur diskriminatif.

Dalam konteks media sosial, kedua jenis ujaran ini sering kali tumpang tindih. Banyak ujaran toksik yang mengandung unsur kebencian, dan sebaliknya, ujaran kebencian sering kali disampaikan dengan gaya bahasa yang toksik. Namun demikian, ujaran toksik memiliki cakupan yang lebih luas, mencakup bentuk ujaran yang ofensif, sinis, atau sarkastik meskipun tidak selalu mengandung kebencian ideologis. (Pavlopoulos dkk., 2020) menegaskan bahwa model deteksi ujaran toksik perlu mampu mengenali spektrum toksisitas mulai dari ujaran kasar ringan hingga ujaran berpotensi diskriminatif, karena dampaknya sama-sama dapat mengganggu ruang komunikasi publik yang sehat.

Secara global, fenomena ujaran toksik semakin menonjol seiring meningkatnya aktivitas masyarakat di media sosial, terutama selama periode diskursus publik yang intens, seperti debat politik atau peristiwa sosial yang memecah belah. Media sosial seperti Twitter (X) menjadi ruang publik digital yang sangat dinamis, di mana opini sering kali disampaikan dengan emosi tinggi, bahkan

disertai serangan verbal terhadap pihak tertentu (Novry dkk., 2025). Kehadiran ujaran toksik dalam diskursus publik tidak hanya mengancam kualitas dialog digital, tetapi juga dapat memperkuat polarisasi masyarakat. Konten yang bersifat ofensif berpotensi memicu konflik sosial, menyebarkan kebencian, dan menciptakan lingkungan komunikasi yang tidak sehat (LOPEZ dkk., 2024).

Tantangan yang muncul dalam konteks ini adalah bagaimana sistem kecerdasan buatan dapat mengenali ujaran toksik yang sering kali tidak disampaikan secara langsung. Banyak ujaran toksik muncul dalam bentuk sindiran, sarkasme, atau permainan kata yang kontekstual terhadap isu tertentu. Hal ini membuat pendekatan berbasis pencocokan kata kunci tidak lagi efektif, karena model harus mampu memahami makna tersirat serta konteks sosial dan linguistik yang melatarinya. Upaya deteksi atau klasifikasi ujaran toksik memerlukan pendekatan *Natural Language Processing* (NLP) yang lebih adaptif terhadap karakteristik bahasa media sosial. Bahasa informal, singkatan, dan campuran bahasa (seperti *code-mixing*) membuat model konvensional sulit menangkap makna sebenarnya (Sahu & S R, 2024).

Oleh karena itu, penelitian ini lebih memfokuskan pada ujaran toksik dibandingkan ujaran kebencian, karena toksisitas mencakup spektrum yang lebih luas dari ekspresi negatif yang muncul dalam wacana daring (Kusuma & Rahmawati, 2025). Dengan pendekatan yang lebih kontekstual melalui model *transformer* seperti RoBERTa, sistem klasifikasi diharapkan mampu menilai toksisitas tidak hanya berdasarkan kata yang digunakan, tetapi juga berdasarkan makna dan intensi ujaran tersebut dalam konteks yang kompleks.

2.4 Klasifikasi Multi-Label dan Pendekatan Hierarkis

Dalam sistem klasifikasi teks, setiap data umumnya dikategorikan ke dalam satu label tertentu. Namun, pada kasus ujaran toksik, satu teks sering kali mengandung lebih dari satu bentuk toksisitas sekaligus. Misalnya, sebuah unggahan di Twitter dapat bersifat *toxic speech* sekaligus *insult* atau *abusive*. Situasi ini menjadikan pendekatan klasifikasi tunggal (*single-label classification*) tidak memadai, sehingga diperlukan pendekatan *multi-label classification* di mana satu entri teks dapat diklasifikasikan ke dalam beberapa kategori toksisitas sekaligus.

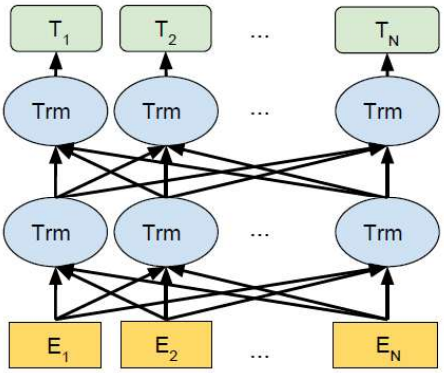
Pendekatan *multi-label* memungkinkan model memahami kompleksitas emosi dan konteks sosial dalam ujaran, tetapi juga menimbulkan tantangan tersendiri. Tantangan tersebut antara lain adalah ketidakseimbangan data antar label (beberapa kategori toksisitas muncul jauh lebih banyak daripada yang lain) serta hubungan antarlabel yang saling berkaitan. Untuk mengatasi hal tersebut, dikembangkan pendekatan *hierarchical multi-label classification*, di mana proses klasifikasi dilakukan secara bertahap. Pada level pertama, sistem menentukan apakah teks termasuk *toxic* atau *non-toxic* (klasifikasi biner). Jika teks terdeteksi sebagai *toxic*, maka dilanjutkan ke level kedua untuk menentukan jenis toksisitas spesifik yang terkandung di dalamnya, meliputi *polarized, profanity_obscenity, threat_incitement_to_violence, insults, identity_attack, sexually_explicit*.

Pendekatan hierarkis ini memberikan beberapa keuntungan. Pertama, model menjadi lebih fokus karena setiap level mempelajari permasalahan yang lebih spesifik. Kedua, pendekatan ini dapat mengurangi kesalahan klasifikasi dengan cara memisahkan antara tahap deteksi toksisitas umum dan identifikasi tipe toksisitas

yang lebih rinci. Ketiga, model hierarkis lebih mudah dievaluasi karena dapat menggunakan metrik berbeda pada setiap tingkat klasifikasi. Dalam konteks penelitian ini, struktur label bertingkat seperti yang digunakan dalam *IndoDiscourse* sangat cocok diolah dengan pendekatan hierarkis, mengingat setiap teks dapat memiliki satu atau lebih kategori toksisitas. Dengan demikian, pendekatan klasifikasi hierarkis ujaran toksik multi-label menjadi strategi yang logis untuk mengatasi kompleksitas dan heterogenitas ujaran berbahasa Indonesia di media sosial.

2.5 Robustly Optimized BERT Approach

RoBERTa (Robustly Optimized BERT Pre-training Approach) adalah model yang dikembangkan dengan mengoptimalkan *BERT* (Bidirectional Encoder Representations from Transformers). Secara struktural, *RoBERTa* menggunakan arsitektur yang serupa dengan *BERT*, yang didasarkan pada Transformer encoder yang terdiri dari beberapa lapisan blok encoder berurutan(Wang dkk., 2023). Gambar 2.1 menggambarkan secara jelas bagaimana setiap token dalam teks yang diberikan akan diproses melalui serangkaian *multi-head self-attention* dan *feed-forward neural network (FFN)* dalam setiap blok encoder untuk menghasilkan representasi kontekstual.



Gambar 2.1 Arsitektur model RoBERTa. Sumber: (Wang dkk., 2023)

Pada gambar 2.1, setiap token yang masuk pertama kali diproses melalui embedding layer, yang mengonversi token menjadi vektor numerik yang mewakili kata-kata dalam kalimat. Setelah melalui beberapa blok *Transformer*, representasi akhir dari token khusus [CLS] akan digunakan sebagai representasi keseluruhan kalimat. Proses ini memungkinkan model untuk menangkap hubungan semantik antar token secara dua arah (*bidirectional*), yang membedakannya dari model sekuensial seperti *RNN* atau *LSTM*. Setiap blok dalam *Transformer* menggunakan self-attention untuk memberikan perhatian terhadap token-token lain dalam kalimat, dan *feed-forward network* yang mengubah hasil atensi menjadi representasi yang lebih kompleks. Dengan demikian, *RoBERTa* mampu memahami konteks lebih dalam untuk setiap token dengan memperhatikan seluruh urutan kalimat secara bersamaan.

Walaupun arsitektur dasar *RoBERTa* menggunakan *Transformer* yang sama dengan *BERT*, perbedaan utama terletak pada tahap *pre-training* dan pengelolaan data. *RoBERTa* mengoptimalkan beberapa aspek pelatihan untuk menghasilkan performa yang lebih baik, terutama dalam menangani teks yang lebih kompleks dan

beragam. Berdasarkan Tabel 2.2, perbedaan utama antara *BERT* dan *RoBERTa* terletak pada tiga aspek utama: tugas *pre-training*, strategi masking, dan ukuran korpus pelatihan yang digunakan.

Tabel 2.2 Perbedaan BERT dan RoBERTa.

Aspek	**BERT**	**RoBERTa**
Tugas Pre-training	MLM dan NSP	MLM
Masking Strategy	*Static*	*Dynamic*
Korpus Pelatihan	±16 GB (BooksCorpus + Wikipedia)	±160 GB (CommonCrawl News, OpenWebText, Stories, CC-News)

Meskipun *BERT* dan *RoBERTa* memiliki struktur arsitektur yang sama, perbedaan utama terletak pada tugas *pre-training* yang digunakan, di mana *RoBERTa* menghapus *Next Sentence Prediction (NSP)* dan hanya menggunakan *Masked Language Modeling (MLM)*. Selain itu, *RoBERTa* menerapkan dynamic masking, di mana token yang disembunyikan dipilih acak setiap *epoch*, memberikan variasi konteks yang lebih banyak. *RoBERTa* juga dilatih dengan korpus yang lebih besar, sekitar 160 GB data dari berbagai sumber yang memungkinkan model memahami variasi linguistik yang lebih beragam. Dengan perubahan-perubahan tersebut, *RoBERTa* menunjukkan performa yang lebih baik dibandingkan *BERT*, terutama pada tugas yang memerlukan pemahaman konteks yang kompleks, seperti deteksi ujaran toksik di media sosial yang memiliki variasi gaya bahasa yang tinggi.

BAB III

DESAIN DAN IMPLEMENTASI SISTEM

3.1 Dataset

Penelitian ini termasuk dalam kategori penelitian terapan (*applied research*) karena berfokus pada penerapan teori dan metode ilmiah dalam pengembangan sistem rekomendasi wisata berbasis teknologi. Hasil penelitian ini diharapkan dapat memberikan solusi praktis bagi wisatawan dalam menentukan destinasi wisata yang sesuai dengan preferensi masing-masing. Penelitian ini menggunakan pendekatan kuantitatif dengan desain eksperimen, karena melibatkan pengumpulan data empiris, penerapan metode *Content-Based Filtering*, serta evaluasi terhadap sistem rekomendasi yang dihasilkan.

Penelitian ini tidak melakukan pengumpulan data primer, melainkan menggunakan dataset yang sudah ada. Tahap ini berfokus pada persiapan data, yang mencakup pemahaman sumber data, analisis karakteristik, dan pemilihan kolom yang relevan. Sumber data penelitian ini menggunakan dataset IndoDiscourse. Dataset ini dikembangkan oleh Susanto et al. (2025) untuk mengatasi kelangkaan data beranotasi multi-label untuk analisis toksisitas dan polarisasi dalam bahasa Indonesia. Data ini dikumpulkan dari berbagai platform media sosial, terutama X (sebelumnya Twitter) dan dapat diakses secara publik melalui repositori Hugging Face (https://huggingface.co/datasets/Exqrch/IndoDiscourse)(Susanto, Wijanarko, Pratama, Tang, dkk., 2025).

Dataset IndoDiscourse yang digunakan dalam penelitian ini terdiri atas 14 atribut, yaitu text_id, annotators_id, text, initial_paragraph, topic,

is_noise_or_spam_text, related_to_election_2024, toxicity, polarized, profanity_obscenity, threat_incitement_to_violence, insults, identity_attack, dan sexually_explicit. Atribut-atribut tersebut merepresentasikan informasi identitas data, hasil anotasi, konteks topik, serta label toksisitas yang digunakan untuk menganalisis ujaran toksik pada media sosial.

Meskipun dataset menyediakan berbagai atribut, tidak seluruh kolom digunakan dalam penelitian ini. Kolom yang dimanfaatkan secara langsung hanya kolom yang berperan sebagai masukan (*input*) model dan label klasifikasi. Sementara itu, atribut seperti *text_id*, *annotators_id*, *initial_paragraph*, *topic*, *is_noise_or_spam_text*, dan *related_to_election_2024* tidak digunakan dalam proses pelatihan maupun evaluasi model karena tidak berkaitan secara langsung dengan tujuan klasifikasi ujaran toksik yang dilakukan. Rincian atribut yang digunakan dalam penelitian ini ditunjukkan pada Tabel 3.1.

Tabel 3.1 Atribut Dataset yang Digunakan dalam Penelitian

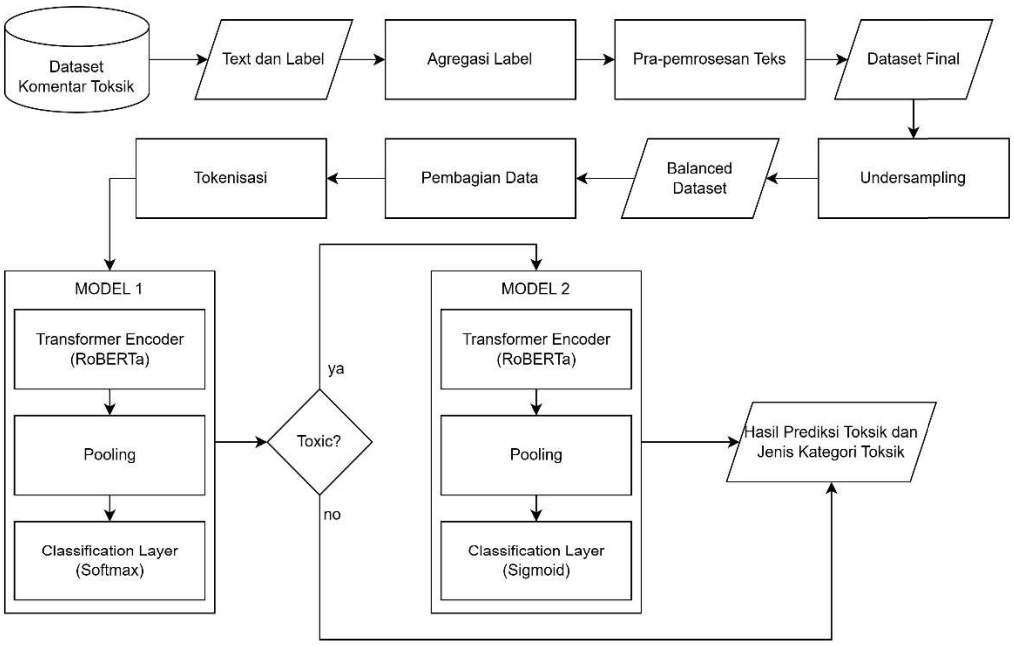
No	Nama Kolom	Deskripsi	Tipe Data
1	text	Teks komentar yang digunakan sebagai masukan utama model.	String
2	toxicity	Label untuk teks mengandung ujaran yang kasar, tidak sopan, tidak masuk akal, penuh kebencian, atau agresif sehingga dapat menimbulkan perasaan tidak nyaman, marah, sedih, terhina, atau membuat seseorang enggan melanjutkan diskusi.	Integer 0/1
3	polarized	Label untuk teks bertujuan mempromosikan konflik antara dua kelompok atau lebih melalui sudut pandang yang sangat bias atau ekstrem terhadap suatu isu tertentu.	Integer 0/1
4	profanity_obscenity	Label untuk teks mengandung kata-kata kasar, vulgar, cabul, atau konten yang dianggap tidak pantas dan bertentangan dengan norma sosial yang berlaku.	Integer 0/1
5	threat_incitement_to_violence	Label untuk teks mengandung ancaman, intimidasi, atau ajakan untuk melakukan kekerasan yang dapat menimbulkan rasa	Integer 0/1

		takut, bahaya, atau tekanan terhadap individu maupun kelompok tertentu.	
6	insults	Label untuk teks mengandung bahasa yang menghina, merendahkan, atau menyinggung dengan tujuan untuk menyakiti perasaan atau meremehkan individu maupun kelompok tertentu.	Integer 0/1
7	identity_attack	Label untuk teks menyerang identitas atau karakteristik seseorang maupun kelompok tertentu, seperti ras, agama, gender, orientasi seksual, penampilan, atau atribut pribadi lainnya.	Integer 0/1
8	sexually_explicit	Label untuk teks mengandung deskripsi atau pembahasan yang eksplisit mengenai aktivitas seksual, bagian tubuh seksual, atau konten seksual lainnya.	Integer 0/1

Dalam penelitian ini, kolom *text* digunakan sebagai masukan utama pada model RoBERTa. Selanjutnya, kolom *toxicity* digunakan sebagai label utama untuk menentukan apakah suatu teks termasuk kategori *toxic* atau *non-toxic*. Apabila suatu teks diklasifikasikan sebagai *toxic*, proses klasifikasi dilanjutkan ke menentukan jenis toksisitas menggunakan enam label kategori ujaran toksik, yaitu *polarized*, *profanity_obscenity*, *threat_incitement_to_violence*, *insults*, *identity_attack*, dan *sexually_explicit*. Struktur label tersebut mendukung penerapan pendekatan klasifikasi hierarkis yang digunakan dalam penelitian ini.

3.2 Desain Sistem

Desain sistem menjelaskan bagaimana proses klasifikasi dilakukan oleh sistem mulai dari data masukan hingga menghasilkan keluaran berupa kategori toksik. Sistem ini menggunakan model *XLM-RoBERTa-base* yang diimplementasikan untuk klasifikasi multi-label pada teks berbahasa Indonesia. Proses kerja sistem ditunjukkan secara umum pada Gambar 3.8 berikut.



Gambar 3.1 Diagram Alur Desain Sistem

Sistem ini menerima input berupa teks komentar atau pernyataan dari dataset yang sudah disiapkan sebelumnya. Setiap teks akan melalui tahap *preprocessing* untuk diubah menjadi format numerik yang dapat dipahami oleh model *transformer*. Selanjutnya, representasi vektor teks tersebut diproses oleh *encoder* dari *XLM-RoBERTa-base* untuk mengekstraksi fitur semantik. Hasil dari proses tersebut diteruskan ke *classification layer* yang berfungsi untuk memprediksi probabilitas pada masing-masing kategori toksik seperti *polarized*, *profanity/obscenity*, *threat/incitement to violence*, *insults*, *identity attack*, dan *sexually explicit*. Output akhir berupa nilai prediksi dari setiap label dalam rentang 0–1.

3.2.1 Agregasi Label

Dataset yang digunakan merupakan hasil anotasi dari beberapa anotator untuk setiap sampel teks. Oleh karena itu, nilai pada setiap label masih berupa kumpulan hasil anotasi, misalnya [0, 1, 1], yang menunjukkan keputusan masing-masing anotator terhadap suatu label. Untuk memperoleh satu label akhir yang dapat digunakan dalam proses pelatihan model, dilakukan proses agregasi menggunakan metode majority voting. Pada metode ini, nilai rata-rata dari seluruh anotasi pada setiap label dihitung terlebih dahulu. Jika nilai rata-rata lebih besar dari 0,5 maka label akhir ditetapkan sebagai 1, sedangkan jika nilai rata-rata lebih kecil dari 0,5 maka label akhir ditetapkan sebagai 0. Sebagai contoh, hasil anotasi [1, 1, 0] memiliki nilai rata-rata 0,67 sehingga dikonversi menjadi label 1. Sebaliknya, hasil anotasi [0, 0, 1] memiliki nilai rata-rata 0,33 sehingga dikonversi menjadi label 0.

Proses agregasi dilakukan terhadap seluruh label dalam dataset, baik label utama (*toxicity*) maupun seluruh label kategori toksik. Setelah nilai rata-rata dihitung, dilakukan pemeriksaan terhadap kemungkinan terjadinya nilai rata-rata 0.5 yang menunjukkan bahwa jumlah anotator yang memberikan label positif dan negatif sama banyak. Kondisi tersebut mengindikasikan tidak adanya keputusan mayoritas sehingga label dianggap ambigu. Pada penelitian ini, apabila terdapat minimal satu label pada suatu sampel yang memiliki nilai rata-rata 0.5, maka seluruh sampel tersebut dihapus dari dataset. Kebijakan ini diterapkan untuk memastikan bahwa setiap sampel yang digunakan memiliki hasil anotasi yang jelas dan konsisten pada seluruh label.

3.2.2 Pra-pemrosesan Teks

a) Menghapus Teks Duplikat

Proses ini dilakukan untuk menghilangkan sampel data yang memiliki kemiripan teks secara identik di dalam dataset. Penghapusan teks duplikat sangat penting untuk mencegah terjadinya kebocoran data (*data leakage*) antara data latih dan data uji, serta menghindari bias pada model yang dapat berujung pada *overfitting*. Dengan memastikan setiap sampel bersifat unik, evaluasi performa model akan menjadi lebih objektif dan representatif.

b) Menghapus Data Non-toksik Tetapi Memiliki Label Jenis Kategori Toksik

Langkah pembersihan ini bertujuan untuk mengatasi anomali atau kontradiksi logis pada hasil anotasi dataset. Secara konseptual, sebuah teks yang dilabeli sebagai non-toksik seharusnya tidak memiliki label turunan apa pun pada label kategori toksik. Data yang mengandung kontradiksi ini dianggap sebagai *noise* akibat kesalahan anotator (*human error*) dan perlu dihapus agar tidak membingungkan model saat proses inferensi hierarkis.

c) Menghapus Data Toksik Tetapi Tidak Ada Label Jenis Kategori Toksik

Serupa dengan poin sebelumnya, langkah ini menangani data yang bernilai tidak konsisten. Teks yang telah diidentifikasi positif sebagai ujaran toksik pada label utama secara logis harus masuk ke dalam minimal satu jenis kategori toksik yang tersedia. Jika terdapat data toksik namun seluruh label kategorinya bernilai

nol (kosong), data tersebut dianggap tidak lengkap (*incomplete annotation*) dan dihapus demi menjaga integritas arsitektur klasifikasi hierarkis.

d) *Case Folding*

proses mengubah seluruh huruf menjadi huruf kecil. Langkah ini diperlukan karena dalam dataset masih banyak teks dengan huruf kapital yang tidak konsisten, sehingga berpotensi dianggap berbeda oleh sistem meskipun memiliki makna yang sama. Seperti pada Tabel 3.2 yang memperlihatkan perbandingan dari proses sebelum dan sesudah *case folding*.

Tabel 3.2 Proses *Case Folding*

Sebelum *Case Folding*	**Setelah *Case Folding***
Pendukung2 boneka Nasdem antek cina!!	pendukung2 boneka nasdem antek cina!!

e) *Numbering and Special Char Removal*

Proses ini menghapus karakter yang tidak relevan seperti URL, simbol, angka, tanda baca berlebih, *emoji*, atau karakter non-alfabet yang tidak memiliki kontribusi terhadap makna teks. Tujuannya agar teks yang dimasukkan ke model bersih dan fokus pada konten semantik utama. Pada contoh teks "Pendukung2 boneka Nasdem antek cina!!" terdapat elemen tambahan yang perlu dihapus yakni berupa tanda seru. Kemudian hasilnya berubah dari teks awal menjadi "pendukung boneka nasdem antek cina".

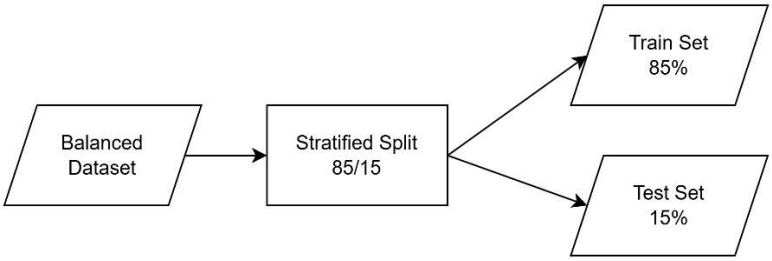
3.2.3 Undersampling

Dataset klasifikasi teks, khususnya pada kasus deteksi ujaran toksik, umumnya memiliki masalah ketidakseimbangan kelas (*class imbalance*) yang signifikan, di mana jumlah komentar non-toksik (kelas mayoritas) jauh lebih

banyak dibandingkan komentar toksik (kelas minoritas). Untuk mengatasi hal ini, diterapkan teknik *undersampling* yang bertujuan mereduksi jumlah sampel pada kelas mayoritas secara acak hingga jumlahnya seimbang dengan kelas minoritas. Proses ini menghasilkan *Balanced Dataset* yang mencegah model cenderung memprediksi kelas mayoritas secara terus-menerus (*bias* kelas mayoritas), sehingga model mampu mengenali pola linguistik dari kelas minoritas dengan lebih optimal.

3.2.3 Pembagian Data

Berbeda dengan pembagian statis sederhana, penelitian ini menerapkan skema *Stratified k-fold cross validation* untuk memastikan evaluasi *model* yang lebih *robust* dan mengurangi bias yang mungkin timbul akibat pemilihan data latih yang spesifik. Alur skema pembagian data yang digunakan ditunjukkan pada Gambar 3.4.



Gambar 3.2 Skema Pembagian Data.

Berdasarkan Gambar 3.4, dataset awal terlebih dahulu dibagi menjadi dua bagian utama menggunakan metode pembagian berstrata (stratified split), yaitu data pengujian (test set) sebesar 15% dan data latihan (training set) sebesar 85%. Pemilihan proporsi ini didasarkan pada pertimbangan metodologis bahwa *test set* harus disediakan sebagai evaluasi akhir yang independen tanpa mengurangi

kecukupan data untuk kebutuhan pengembangan *model*. Literatur pendukung menegaskan pentingnya penggunaan *blind test set* yang dipisahkan secara total dari proses pelatihan untuk menjamin estimasi performa *model* yang lebih andal dan konsisten(Allgaier & Pryss, 2024; Y. Xu & Goodacre, 2018). Dengan demikian, penggunaan 85% data sebagai *training set* dipandang rasional karena mampu memaksimalkan data untuk pelatihan dan validasi silang sekaligus tetap menjaga kemandirian *hold-out test set*.

Data pengujian tersebut sejak awal disisihkan sebagai *hold-out set* dan tidak pernah dilibatkan dalam proses pelatihan maupun validasi silang agar dapat berfungsi sebagai simulasi evaluasi akhir pada data dunia nyata. Selanjutnya, pada tahap *Cross validation*, data pengembangan tersebut dibagi menjadi K bagian atau lipatan (*folds*) yang sama besar dengan nilai K=5 yang telah ditetapkan. Penerapan teknik berstrata ini menjamin bahwa setiap lipatan data memiliki proporsi distribusi kelas toksik yang identik dengan dataset aslinya. Proses pelatihan kemudian dilakukan secara berulang sebanyak lima kali, di mana pada setiap iterasinya terdapat empat bagian data yang digunakan untuk memperbarui bobot model dan satu bagian sisanya digunakan sebagai data validasi guna menghindari masalah penghafalan data atau *overfitting*.

Penggunaan *K-Fold Cross validation* merupakan standar validasi yang lebih ketat dibandingkan pembagian tunggal, sebagaimana disarankan oleh (Wong & Yeh, 2020) untuk mendapatkan estimasi akurasi yang lebih andal. Secara matematis, performa akhir model divalidasi dengan merata-ratakan hasil evaluasi dari kelima iterasi tersebut. Pendekatan ini berfungsi untuk meminimalkan risiko

variasi hasil yang disebabkan oleh faktor kebetulan saat pengambilan sampel secara acak, sehingga metrik evaluasi yang dihasilkan lebih merepresentasikan kemampuan generalisasi model yang sebenarnya sebelum dievaluasi secara final pada data pengujian.

3.2.4 Tokenisasi

Proses tokenisasi menggunakan SentencePiece dengan algoritma *Unigram Language Model* (*Unigram LM*) yang digunakan oleh tokenizer *XLM-RoBERTa-base*. Proses ini mengubah teks menjadi serangkaian token atau subword serta menambahkan token khusus <s> sebagai penanda awal kalimat dan </s> sebagai penanda akhir kalimat. Sebagai ilustrasi, kalimat "pendukung boneka nasdem antek cina" dapat direpresentasikan menjadi token-token seperti <s>, ▁pendukung, ▁boneka, ▁nasdem, ▁antek, ▁cina, dan </s>. Simbol ▁ menunjukkan awal sebuah kata pada tokenizer *SentencePiece*.

3.2.5 Transformer Encoder (RoBERTa)

Setiap token kemudian diubah menjadi representasi numerik (vektor) menggunakan *embedding layer* dari *RoBERTa*, di mana setiap kata diwakili oleh vektor berdimensi tetap (*d_model*). Secara matematis, proses *embedding* dapat dinyatakan pada persamaan 3.1:

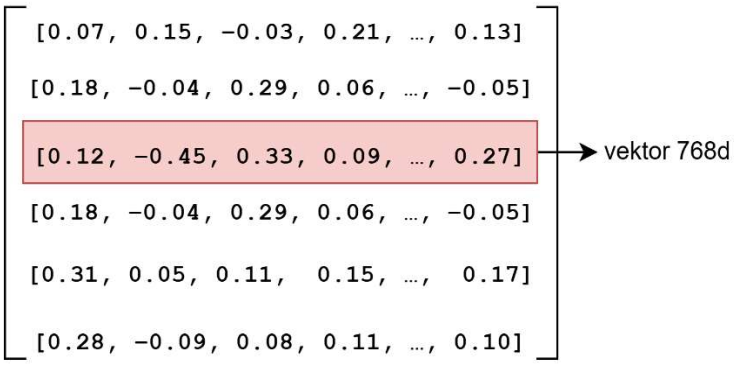
$$E_i = W_e[T_i]$$
$$H_0 = \text{LayerNorm}(E_i + P_i) \quad (3.1)$$

Keterangan:
E_i = vektor token embedding ke-i
W_e = matriks bobot embedding
T_i = indeks token ke-i
P_i = position embedding

Pada tahap ini, representasi teks berubah dari bentuk string menjadi representasi numerik melalui proses *embedding*. Setiap token hasil tokenisasi dikonversi menjadi sebuah vektor numerik dengan ukuran 768 dimensi yang merepresentasikan karakteristik semantik dari token tersebut. Karena satu kalimat terdiri dari beberapa token, seluruh vektor token tersebut digabungkan membentuk sebuah tensor embedding dengan tiga dimensi, yaitu *(batch size, sequence length, hidden dimension).*

Pada penelitian ini, dimensi tensor embedding dapat dituliskan sebagai $(B, L, 768)$ dengan: B *(batch size)* menunjukkan jumlah sampel atau kalimat yang diproses secara bersamaan dalam satu batch. L *(sequence length)* menunjukkan jumlah token dalam satu kalimat setelah proses tokenisasi. Nilai ini mengikuti panjang maksimum token yang digunakan, yaitu hingga 512 token. *768 (hidden dimension)* merupakan ukuran representasi vektor dari setiap token yang dihasilkan oleh model Transformer.

Sebagai contoh, apabila satu kalimat setelah tokenisasi memiliki 10 token dan diproses dalam satu batch, maka bentuk tensor embedding yang dihasilkan adalah $(1, 10, 768)$. Artinya, terdapat 1 kalimat yang direpresentasikan oleh 10 token, dan setiap token memiliki representasi berupa vektor dengan panjang 768 nilai numerik. Tensor inilah yang kemudian menjadi masukan bagi encoder Transformer untuk menangkap hubungan kontekstual antar token. Gambar 3.10 menunjukkan contoh bentuk representasi embedding dari setiap token setelah proses tokenisasi dan embedding.



Gambar 3.3 Bentuk tensor hasil proses tokenization dan embedding

Pada tahap ini, representasi teks berubah dari bentuk string menjadi matriks numerik yang merepresentasikan seluruh token dalam satu kalimat. Keluaran tahap ini adalah tensor embedding dengan dimensi vektor sebanyak 768.

Model *RoBERTa* merupakan varian dari arsitektur *Bidirectional Encoder Representations from Transformers (BERT)* yang ditingkatkan dengan pelatihan yang lebih intensif dan penyesuaian pada *training objective*. Pada penelitian ini digunakan versi *base multilingual* agar dapat memahami konteks lintas bahasa, termasuk bahasa Indonesia. Setiap input token dari hasil *embedding* diproses melalui beberapa *encoder layer* yang masing-masing terdiri dari dua komponen utama, yaitu *multi-head self-attention* dan *feed-forward network*.

a) Self-Attention

Mekanisme *self-attention* merupakan komponen utama dalam arsitektur Transformer yang memungkinkan model memahami hubungan antar kata dalam satu kalimat secara kontekstual. *Self-attention* memungkinkan setiap token memperhatikan seluruh token lain dalam satu kalimat secara simultan. Dengan

mekanisme ini, model dapat menangkap ketergantungan jarak dekat maupun jarak jauh antar kata.

Pada tahap awal, setiap token hasil *embedding* direpresentasikan dalam bentuk vektor dan disusun menjadi sebuah matriks *embedding* input X. Matriks ini kemudian ditransformasikan secara linear menjadi tiga representasi berbeda, yaitu query (Q), key (K), dan value (V). Transformasi ini dinyatakan pada persamaan 3.2:

$$Q = XW^Q, K = XW^K, V = XW^V \quad (3.2)$$

Keterangan:
X = matriks representasi token hasil proses *embedding*.
Q = matriks *Query* untuk mencari relevansi antar token.
K = matriks *Key* sebagai acuan pencocokan perhatian (*attention*).
V = matriks *Value* yang menyimpan informasi yang akan diteruskan ke lapisan berikutnya.
W^Q = matriks bobot *Query*.
W^K = matriks bobot *Key*.
W^V = matriks bobot *Value*.

Pada persamaan tersebut, X merupakan matriks embedding input berukuran n × d, dengan n menyatakan jumlah token dan d menyatakan dimensi *embedding*. Matriks W^Q, W^K, dan W^V merupakan matriks bobot yang dipelajari selama proses *fine-tuning*. Hasil dari transformasi ini adalah tiga matriks Q, K, dan V yang masing-masing merepresentasikan sudut pandang berbeda dari token yang sama.

Selanjutnya, mekanisme *self-attention* menghitung seberapa besar perhatian suatu token terhadap token lain dalam satu kalimat. Nilai perhatian dihitung dengan mengalikan matriks Q dengan transpose dari matriks K, kemudian dinormalisasi menggunakan fungsi *softmax*. Proses ini dirumuskan pada persamaan 3.3:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (3.3)$$

Keterangan:

$Attention(Q, K, V)$	= keluaran mekanisme *self-attention*.
Q	= matriks *Query*.
K	= matriks *Key*.
V	= matriks *Value*.
K^T	= transpos dari matriks *Key*.
QK^T	= skor kesamaan (*attention score*) antara setiap pasangan token.
d_k	= dimensi vektor *Key*.
$\sqrt{d_k}$	= faktor normalisasi untuk menjaga stabilitas nilai *softmax*.
$softmax(\cdot)$	= fungsi yang mengubah skor perhatian menjadi distribusi probabilitas.

Pada persamaan 3.3, d_k merupakan dimensi vektor key yang digunakan sebagai faktor penskalaan untuk menjaga stabilitas nilai gradien. Operasi QK^T menghasilkan matriks skor perhatian yang menunjukkan tingkat relevansi antar token. Fungsi *softmax* kemudian mengubah skor tersebut menjadi distribusi probabilitas, sehingga jumlah bobot perhatian untuk setiap token bernilai satu. Distribusi ini digunakan untuk melakukan kombinasi berbobot terhadap matriks V, yang menghasilkan representasi baru bagi setiap token.

Sebagai ilustrasi konseptual, pada kalimat “Pendukung boneka Nasdem antek cina”, token “cina” akan memiliki nilai perhatian yang tinggi terhadap token “antek” dan “Pendukung”. Hal ini terjadi karena secara semantik, kata-kata tersebut saling berhubungan dalam konteks serangan terhadap kelompok tertentu. Dengan demikian, representasi token “cina” tidak hanya bergantung pada makna katanya sendiri, tetapi juga pada konteks keseluruhan kalimat.

Untuk meningkatkan kapasitas representasi, RoBERTa tidak hanya menggunakan satu mekanisme *self-attention*, tetapi menerapkan *multi-head attention*. Pendekatan ini memungkinkan model mempelajari berbagai jenis

hubungan semantik secara paralel. Proses *multi-head attention* dirumuskan pada persamaan 3.4:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (3.4)$$

Keterangan:
$MultiHead(Q, K, V)$ = keluaran mekanisme *Multi-Head Attention*.
$head_1, \dots, head_h$ = hasil perhatian dari masing-masing *attention head*.
h = jumlah *attention head* yang digunakan.
$Concat(\cdot)$ = operasi penggabungan keluaran seluruh *head*.
W^O = matriks bobot keluaran (*output projection matrix*).

Setiap head merupakan hasil dari mekanisme *self-attention* yang dihitung secara independen dengan bobot yang berbeda, sebagaimana dinyatakan pada persamaan 3.5:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (3.5)$$

Keterangan:
$head_i$ = keluaran *attention head* ke-i.
Q, K, dan V = matriks *Query*, *Key*, dan *Value*.
W_i^Q = matriks bobot *Query* pada *head* ke-i.
W_i^K = matriks bobot *Key* pada *head* ke-i.
W_i^V = matriks bobot *Value* pada *head* ke-i.
$Attention(\cdot)$ = fungsi *Scaled Dot-Product Attention* pada Persamaan (3.3).
i = indeks attention head, dengan $i = 1,2, \dots, h$.

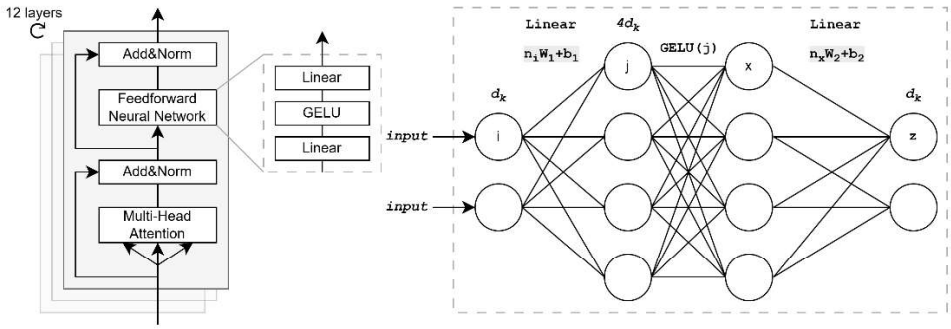
Dalam penelitian ini, jumlah head yang digunakan adalah 12, sesuai dengan konfigurasi RoBERTa-base. Matriks W_i^Q, W_i^K, dan W_i^V merupakan bobot transformasi untuk setiap head, sedangkan W^O merupakan matriks bobot output yang menggabungkan seluruh head menjadi satu representasi. Hasil akhir dari *multi-head attention* adalah representasi kontekstual token yang lebih kaya dan mampu menangkap berbagai sudut pandang hubungan antar kata.

3.2.6 Pooling

Lapisan *pooling* digunakan untuk mereduksi dimensi keluaran dari *Transformer Encoder* yang awalnya berupa sekuens vektor (*hidden states*) untuk

setiap token, menjadi satu representasi vektor tunggal (*fixed-size vector*) yang merepresentasikan keseluruhan makna teks. Pada penelitian ini, proses *pooling* dilakukan dengan mengambil *hidden state* dari token pertama pada sekuens <s> yang merepresentasikan seluruh konteks kalimat, lalu melewatkannya pada sebuah lapisan *dense* linier yang dilengkapi dengan fungsi aktivasi. Vektor hasil *pooling* inilah yang menyimpan ringkasan informasi semantik teks.

3.2.7 Classification Layer



Gambar 3.4 Diagram Alur Classification Layer

Keluaran dari *multi-head attention* pada transformer selanjutnya diproses oleh *Feed-forward Neural Network* untuk memberikan transformasi non-linear pada representasi token. *Feed-forward neural network* pada arsitektur Transformer terdiri dari dua lapisan linear dengan fungsi aktivasi GELU di antaranya sesuai Gambar 3.7. Fungsi linear pertama merubah shape menjadi 4x dimensi input dimana pada RoBERTa menggunakan dimensi input sebanyak 768. Setelah itu, masuk ke fungsi aktivasi *GELU* digunakan untuk memperkenalkan non-linearitas sehingga model mampu mempelajari pola yang lebih kompleks. Keluaran dari

GELU akan dibawa lagi ke fungsi linear yang kedua. Dimana ini merubah dimensinya kembali ke awal yakni 768. Proses ini dirumuskan pada persamaan 3.6:

$$\mathrm{FNN}(x) = \mathrm{GELU}(xW_1 + b_1)W_2 + b_2 \quad (3.6)$$

Keterangan:
$FNN(x)$ = keluaran dari *Feedforward Neural Network*.
x = representasi token hasil keluaran lapisan *Multi-Head Attention*.
W_1 = matriks bobot pada lapisan linear pertama.
b_1 = vektor bias pada lapisan linear pertama.
W_2 = matriks bobot pada lapisan linear kedua.
b_2 = vektor bias pada lapisan linear kedua.
$GELU(\cdot)$ = fungsi aktivasi *Gaussian Error Linear Unit*
$xW_1 + b_1$ = transformasi linear pertama yang memproyeksikan representasi token ke dimensi yang lebih tinggi.
$GELU(xW_1 + b_1)$ = hasil aktivasi non-linear yang memungkinkan model mempelajari pola yang lebih kompleks.

Sebagai contoh konseptual, jika keluaran *multi-head attention* menghasilkan vektor representasi yang menekankan hubungan antara kata “antek” dan “cina”, maka *feed-forward network* akan mentransformasikan representasi tersebut ke dalam ruang fitur yang lebih diskriminatif. Proses ini memungkinkan model membedakan konteks ujaran yang bersifat netral dengan konteks yang mengandung serangan identitas atau hinaan. Output dari feed-forward network berupa representasi kontekstual non-linear untuk setiap token input. Representasi ini diteruskan ke lapisan encoder berikutnya hingga mencapai lapisan terakhir. Setelah itu akan dipilih token pertama <*s*> sebagai token representasi. Kemudian akan masuk ke layer softmax atau sigmoid sesuai yang dibutuhkan. Dimana fungsi matematis yang berbeda untuk mengakomodasi kebutuhan spesifik setiap model:

a) *Softmax Layer (Model 1)*

Pada tahap pertama, vektor representasi token <s> yang dihasilkan oleh encoder diproyeksikan ke ruang keluaran menggunakan lapisan *fully connected*. Lapisan ini menghasilkan dua nilai *logit* yang merepresentasikan kelas *non-toxic* dan *toxic* dengan menggunakan persamaan 3.7.

$$z^{(1)} = W_1 h_{\langle s \rangle} + b_1 \tag{3.7}$$

Keterangan:
$z^{(1)}$ = vektor *logit* keluaran Level 1.
$h_{\langle s \rangle} \in \mathbb{R}^d$ = representasi token <s> hasil keluaran *RoBERTa Encoder*.
$W_1 \in \mathbb{R}^{2 \times d}$ = matriks bobot lapisan klasifikasi biner.
$b_1 \in \mathbb{R}^2$ = vektor bias lapisan klasifikasi biner.
d = dimensi *hidden state* RoBERTa.

Karena klasifikasi pada Level 1 merupakan klasifikasi biner, maka nilai *logit* dikonversi menjadi probabilitas menggunakan fungsi *softmax* seperti pada persamaan 3.8:

$$p^{(1)} = softmax(z^{(1)}) \tag{3.8}$$

Keterangan:
$p^{(1)} = [p_{non-toxic}, p_{toxic}]$ = probabilitas untuk masing-masing kelas.
p_{toxic} = probabilitas bahwa teks termasuk kategori *toxic*.
$p_{non-toxic}$ = probabilitas bahwa teks termasuk kategori *non-toxic*.

Pada implementasi penelitian ini, nilai probabilitas kelas toxic digunakan sebagai dasar pengambilan keputusan.

$$\hat{y}^{(1)} = \begin{cases} 1, & \text{jika } p_{toxic} \geq \tau_1 \\ 0, & \text{jika } p_{toxic} < \tau_1 \end{cases} \tag{3.9}$$

Keterangan:
$\hat{y}^{(1)}$ = hasil prediksi Level 1.
1 = kelas *toxic*.
0 = kelas *non-toxic*.
τ_1 = nilai *threshold* Level 1.

Pada penelitian ini digunakan nilai *threshold* optimal hasil evaluasi model.

Apabila suatu teks diprediksi sebagai *non-toxic*, maka proses klasifikasi dihentikan. Sebaliknya, apabila teks diprediksi sebagai *toxic*, maka teks diteruskan ke proses klasifikasi Level 2.

b) *Sigmoid Layer (Model 2)*

Pada Model Level 2, representasi token <s> yang sama digunakan kembali untuk mengidentifikasi kategori toksisitas yang terkandung dalam teks. Berbeda dengan Level 1, tahap ini menggunakan pendekatan *multi-label classification* sehingga satu teks dapat memiliki lebih dari satu label secara bersamaan.

Lapisan klasifikasi menghasilkan sejumlah *logit* sesuai dengan jumlah label yang digunakan dalam penelitian.

$$z^{(2)} = W_c h_{\langle s \rangle} + b_c \tag{3.10}$$

Keterangan:
$z^{(2)} \in \mathbb{R}^L$: vektor *logit* keluaran Level 2.
$W_c \in \mathbb{R}^{L \times d}$: matriks bobot klasifikasi multi-label.
$b_c \in \mathbb{R}^L$: vektor bias klasifikasi multi-label.
L: jumlah label jenis toksisitas yang digunakan pada penelitian.

Nilai *logit* kemudian diubah menjadi probabilitas menggunakan fungsi *sigmoid* yang diterapkan secara independen pada setiap label.

$$\hat{p} = \sigma(z^{(2)}) \tag{3.11}$$

Keterangan:
$\hat{p} = [p_1, p_2, \dots, p_L]$: vektor probabilitas seluruh label.
p_i: probabilitas kemunculan label ke-i.
$\sigma(\cdot)$: fungsi aktivasi *sigmoid*.

Karena menggunakan pendekatan *multi-label classification*, setiap label diprediksi secara independen dengan nilai *threshold* yang berbeda untuk setiap label.

$$\hat{y}_i = \begin{cases} 1, & \text{jika } p_i \geq \tau_i \\ 0, & \text{jika } p_i < \tau_i \end{cases} \quad (3.12)$$

Keterangan:
$\hat{y}_i$ = hasil prediksi label ke-i.
p_i = probabilitas label ke-i.
τ_i = *threshold* optimal untuk label ke-i.
1 menunjukkan label terdeteksi pada teks.
0 menunjukkan label tidak terdeteksi pada teks.

Melalui mekanisme tersebut, satu teks dapat memperoleh lebih dari satu label jenis toksisitas sesuai dengan probabilitas yang dihasilkan oleh model. Pendekatan ini memungkinkan sistem untuk mengidentifikasi berbagai bentuk ujaran toksik yang muncul secara bersamaan dalam satu komentar.

3.3 Lingkungan Penerapan dan Pengujian Sistem

Pengembangan, pelatihan, validasi, dan pengujian model pada penelitian ini dilakukan menggunakan lingkungan komputasi berbasis *cloud* melalui layanan Google Colaboratory (*Google Colab Pro*). Pemanfaatan platform *cloud computing* dipilih karena mampu menyediakan sumber daya komputasi yang memadai untuk menjalankan model *deep learning* tanpa memerlukan infrastruktur perangkat keras lokal dengan spesifikasi tinggi. Selain itu, lingkungan Google Colab menyediakan integrasi langsung dengan berbagai pustaka *machine learning* dan *deep learning* yang mendukung proses eksperimen secara efisien dan terukur.

Pada penelitian ini, proses pelatihan model dilakukan dengan memanfaatkan akselerasi *Graphics Processing Unit* (GPU) NVIDIA Tesla T4 yang

tersedia pada layanan Google Colab Pro. Penggunaan GPU diperlukan untuk mempercepat proses komputasi, khususnya pada tahapan *forward propagation*, *backpropagation*, serta pembaruan parameter model selama proses pelatihan. Spesifikasi lingkungan komputasi yang digunakan dalam penelitian ini disajikan pada Tabel 3.3.

Tabel 3.3 Spesifikasi Lingkungan Komputasi Pengujian

Komponen	Spesifikasi
Platform Komputasi	Google Colaboratory Pro
Jenis Lingkungan	Cloud-based Jupyter Notebook
Akselerator GPU	NVIDIA Tesla T4
Kapasitas VRAM GPU	16 GB GDDR6
Arsitektur GPU	NVIDIA Turing
Memori Sistem (RAM)	±12 GB – 25 GB
Media Penyimpanan	Google Drive

Selain infrastruktur komputasi, keberhasilan proses pelatihan model juga dipengaruhi oleh konfigurasi perangkat lunak yang digunakan. Penelitian ini memanfaatkan bahasa pemrograman Python beserta berbagai pustaka pendukung untuk pengolahan data, representasi teks, pelatihan model transformer, dan evaluasi performa klasifikasi. Rincian perangkat lunak yang digunakan disajikan pada Tabel 3.4.

Tabel 3.4 Spesifikasi Perangkat Lunak Pengujian

Komponen Perangkat Lunak	Spesifikasi / Versi
Sistem Operasi	Linux Ubuntu (Google Colab Runtime)
Bahasa Pemrograman	Python 3.12
Framework Deep Learning	PyTorch 2.11
Library Transformer	Transformers (Hugging Face)
Akselerasi GPU	CUDA 12.8
Pustaka Pemrosesan Data	Pandas 2.2, NumPy 2.0
Pustaka Evaluasi Model	Scikit-Learn
Pustaka Model	XLM-RoBERTa
Lingkungan Pengembangan	Google Colaboratory Pro

Selain spesifikasi perangkat keras dan perangkat lunak, penelitian ini juga menetapkan konfigurasi *hyperparameter* yang digunakan secara konsisten pada seluruh eksperimen. Penggunaan konfigurasi yang seragam bertujuan untuk menjaga keadilan perbandingan performa antar model, meminimalkan variasi yang disebabkan oleh perbedaan parameter pelatihan, serta meningkatkan reproduktibilitas hasil penelitian. Seluruh model dilatih menggunakan konfigurasi *hyperparameter* yang sama pada setiap proses pelatihan dan evaluasi. Rincian konfigurasi yang digunakan ditunjukkan pada Tabel 3.5.

Tabel 3.5 Konfigurasi *Hyperparameter* Pelatihan

Hyperparameter	**Nilai**
Random Seed	42
Train Batch Size	16
Evaluation Batch Size	64
Learning Rate	2×10^{-5}
Maximum Sequence Length	512
Stratified K-Fold Splits	5
Epoch per Fold	3

Konfigurasi *random seed* sebesar 42 digunakan untuk memastikan proses inisialisasi parameter model, pembagian data, serta proses pelatihan menghasilkan keluaran yang konsisten pada setiap eksperimen. Parameter *train batch size* sebesar 16 dipilih untuk menyeimbangkan kebutuhan memori GPU dengan stabilitas proses pembelajaran model. Sementara itu, *evaluation batch size* sebesar 64 digunakan untuk mempercepat proses inferensi dan evaluasi karena tidak memerlukan proses propagasi balik (*backpropagation*).

Nilai *learning rate* sebesar 2×10^{-5}digunakan karena merupakan salah satu konfigurasi yang umum direkomendasikan pada proses *fine-tuning* model berbasis Transformer, termasuk RoBERTa dan IndoRoBERTa. Parameter *maximum sequence length* ditetapkan sebesar 512 token untuk memanfaatkan kapasitas maksimum model dalam merepresentasikan konteks teks secara utuh.

Pada tahap validasi, penelitian ini menerapkan metode *Stratified K-Fold Cross Validation* dengan jumlah lipatan (*fold*) sebanyak 5. Pendekatan ini dipilih untuk menjaga distribusi kelas pada setiap *fold* tetap proporsional terhadap distribusi data asli. Setiap *fold* dilatih selama 3 *epoch*, sehingga model memiliki kesempatan yang cukup untuk mempelajari pola data tanpa meningkatkan risiko *overfitting* secara berlebihan.

Untuk menjamin konsistensi hasil eksperimen, konfigurasi deterministik pada backend PyTorch diaktifkan dan seluruh proses pelatihan dijalankan menggunakan nilai *random seed* yang sama. Dengan demikian, variasi hasil yang disebabkan oleh proses acak dapat diminimalkan sehingga evaluasi performa model menjadi lebih objektif dan dapat direproduksi pada lingkungan komputasi yang setara.

Lingkungan pengujian ini digunakan secara konsisten pada seluruh tahapan penelitian, mulai dari proses *preprocessing* data, pelatihan model klasifikasi hierarkis berbasis RoBERTa, optimasi parameter, hingga evaluasi performa model menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Dengan demikian, hasil eksperimen yang diperoleh dapat direproduksi dan diverifikasi pada lingkungan komputasi yang setara.

BAB IV

UJI COBA DAN PEMBAHASAN

4.1 Desain Uji Coba

Untuk mengevaluasi performa klasifikasi, penelitian ini melakukan pengujian pada tiga skenario yang berbeda. Skenario pertama fokus pada klasifikasi biner, yaitu menentukan apakah suatu teks termasuk *toxic* atau *non-toxic*. Skenario kedua menggunakan pendekatan *flat multilabel*, di mana model memprediksi seluruh label toksisitas secara bersamaan tanpa struktur hierarki. Skenario ketiga menerapkan pendekatan *hierarchical multilabel classification*, di mana Level 1 bertugas menentukan apakah teks mengandung toksisitas atau tidak, dan Level 2 mengidentifikasi jenis toksisitas apabila teks terdeteksi mengandung unsur *toxic*.

4.1.1 Persiapan Data

Dataset yang digunakan merupakan hasil anotasi dari beberapa anotator untuk setiap sampel teks. Oleh karena itu, nilai pada setiap label masih berupa kumpulan hasil anotasi, misalnya [0, 1, 1], yang menunjukkan keputusan masing-masing anotator terhadap suatu label. Untuk memperoleh satu label akhir yang dapat digunakan dalam proses pelatihan model, dilakukan proses agregasi menggunakan metode majority voting. Pada metode ini, nilai rata-rata dari seluruh anotasi pada setiap label dihitung terlebih dahulu. Jika nilai rata-rata lebih besar dari 0,5 maka label akhir ditetapkan sebagai 1, sedangkan jika nilai rata-rata lebih kecil dari 0,5 maka label akhir ditetapkan sebagai 0. Sebagai contoh, hasil anotasi [1, 1, 0] memiliki nilai rata-rata 0,67 sehingga dikonversi menjadi label 1.

Sebaliknya, hasil anotasi [0, 0, 1] memiliki nilai rata-rata 0,33 sehingga dikonversi menjadi label 0.

Proses agregasi dilakukan terhadap seluruh label dalam dataset, baik label utama (*toxicity*) maupun seluruh label kategori toksik. Setelah nilai rata-rata dihitung, dilakukan pemeriksaan terhadap kemungkinan terjadinya nilai rata-rata 0.5 yang menunjukkan bahwa jumlah anotator yang memberikan label positif dan negatif sama banyak. Kondisi tersebut mengindikasikan tidak adanya keputusan mayoritas sehingga label dianggap ambigu. Pada penelitian ini, apabila terdapat minimal satu label pada suatu sampel yang memiliki nilai rata-rata 0.5, maka seluruh sampel tersebut dihapus dari dataset. Kebijakan ini diterapkan untuk memastikan bahwa setiap sampel yang digunakan memiliki hasil anotasi yang jelas dan konsisten pada seluruh label.

Hasil pemeriksaan menunjukkan terdapat 4.036 sampel ambigu yang memiliki nilai rata-rata 0.5 pada setidaknya satu label. Sampel-sampel tersebut dihapus dari dataset sehingga jumlah data berkurang dari 28.448 sampel menjadi 24.412 sampel. Selanjutnya, seluruh nilai label yang tersisa dikonversi menjadi label biner 0 atau 1 berdasarkan hasil majority voting dan digunakan pada tahap pembersihan data berikutnya. Setelah proses agregassi, dilakukan identifikasi dan penghapusan data anomali untuk memastikan konsistensi pelabelan pada dataset. Dataset yang digunakan memiliki satu label utama (*toxicity*) dan beberapa label kategori toksik yang merepresentasikan jenis toksisitas secara lebih spesifik. Tahap pertama adalah penghapusan data duplikat berdasarkan isi teks. Dari total dataset

awal, ditemukan sebanyak 2.062 teks duplikat. Setelah duplikat dihapus, jumlah data menjadi 22.350 sampel.

Karena penelitian ini membangun struktur hierarkis yang mensyaratkan seluruh label kategori toksisitas berada di bawah label utama *toxicity*, sampel dengan ketidaksesuaian antara label *toxicity* dan label kategori dihapus. Keputusan ini merupakan bagian dari rekonstruksi dataset agar sesuai dengan skema *hierarchical classification*. Selanjutnya dilakukan pemeriksaan konsistensi antara label *toxicity* dan label kategori toksik. Secara konseptual, apabila suatu teks memiliki setidaknya satu kategori toksik aktif, maka teks tersebut seharusnya diberi label *toxicity* bernilai 1. Sebaliknya, apabila label *toxicity* bernilai 1, maka setidaknya harus terdapat satu kategori toksik yang aktif untuk menjelaskan jenis toksisitas yang terkandung dalam teks. Berdasarkan pemeriksaan tersebut ditemukan dua jenis anomali:

1. Label toxicity = 0 tetapi memiliki kategori toksik aktif.

Kondisi ini menunjukkan ketidaksesuaian pelabelan karena teks dikategorikan sebagai tidak toksik, namun masih memiliki label kategori toksik. Ditemukan sebanyak 2.302 sampel dengan kondisi tersebut sehingga dihapus dari dataset.

2. Label toxicity = 1 tetapi tidak memiliki kategori toksik aktif.

Kondisi ini juga dianggap tidak konsisten karena teks dikategorikan sebagai toksik, tetapi tidak memiliki kategori yang menjelaskan bentuk toksisitasnya. Ditemukan sebanyak 104 sampel dengan kondisi tersebut dan dihapus dari dataset.

Setelah seluruh proses penghapusan data anomali dilakukan, jumlah data berkurang dari 22.350 sampel menjadi 19.944 sampel. Penghapusan data anomali ini bertujuan untuk meningkatkan kualitas dataset serta memastikan konsistensi hubungan antara label utama dan kategori toksik sebelum digunakan pada proses pelatihan dan evaluasi model.

Ilustrasi Sampel Data Tabel 4.1 di bawah ini menyajikan tiga contoh data yang merepresentasikan karakteristik dataset, mulai dari data non-toksik hingga data toksik yang memiliki satu atau lebih label jenis toksisitas.

Tabel 4.1 Ilustrasi Sampel Data dari Dataset IndoDiscourse

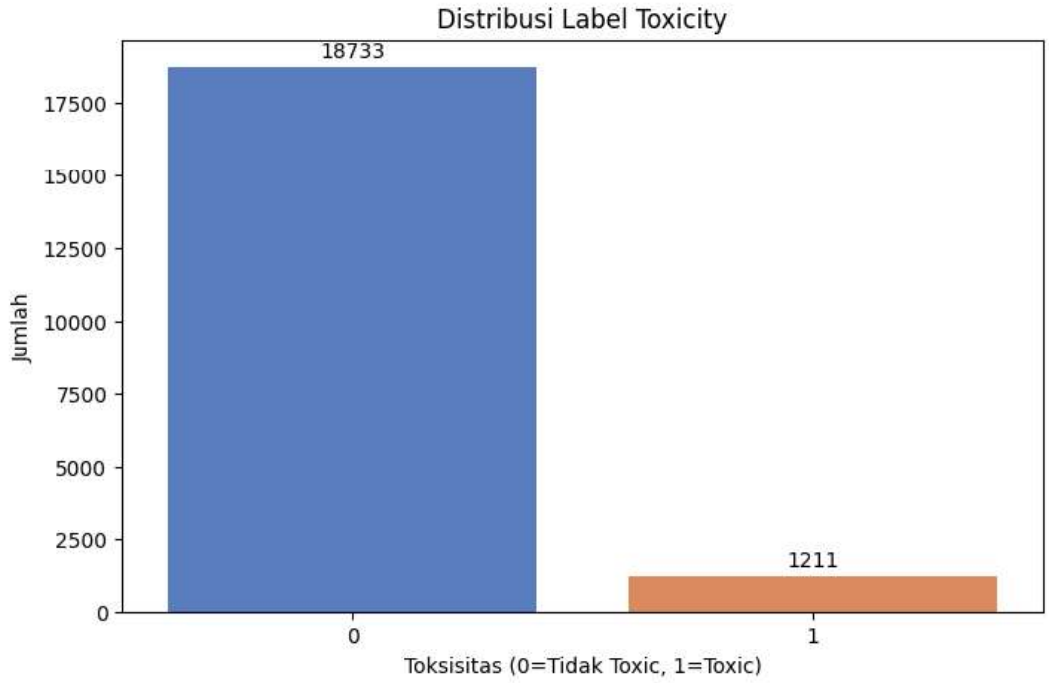
Kategori	**Contoh Teks**	**Toxicity**	**Kategori Ujaran Toksik**					
			1	**2**	**3**	**4**	**5**	**6**
Non-Toxic	happy banget temenku lolos lpdp ke melbourne university gila one shoot doang semoga nular	0	0	0	0	0	0	0
Toxic (Satu Label)	udah enek minta maaf berkali tetep bikin sinting anjgggg	1	0	1	0	0	0	0
Toxic (Multi-Label)	kaum yahudi tanamkan keraguan pada umat islam pagi beriman sorenya murtad	1	1	0	0	0	1	0

Keterangan:
Kolom 1 merepresentasikan label *polarized*,
Kolom 2 merepresentasikan label *profanity_obscenity*,
Kolom 3 merepresentasikan label *threat_incitement_to_violence*,
Kolom 4 merepresentasikan label *insults*,
Kolom 5 merepresentasikan label *identity_attack*,
Kolom 6 merepresentasikan label *sexually_explicit*.

Berdasarkan Tabel 4.1, data *non-toxic* tidak memiliki label aktif pada kategori ujaran toksik sehingga seluruh nilai pada kolom 1–6 bernilai 0. Sebaliknya, data *toxic* dapat memiliki satu maupun lebih dari satu label aktif secara bersamaan. Sebagai contoh, data pada kategori *Toxic (Satu Label)* hanya memiliki label *profanity_obscenity* yang ditunjukkan oleh nilai 1 pada kolom 2. Sementara itu, data pada kategori *Toxic (Multi-Label)* memiliki dua label aktif, yaitu *polarized* dan *identity_attack*, yang ditunjukkan oleh nilai 1 pada kolom 1 dan kolom 5.

Karakteristik ini menunjukkan bahwa dataset yang digunakan memiliki struktur *hierarchical multi-label classification*, di mana satu teks dapat memiliki lebih dari satu kategori ujaran toksik secara bersamaan.

Setelah seluruh proses pembersihan data dilakukan, diperoleh sebanyak 19.944 data yang digunakan dalam penelitian ini. Distribusi label *toxicity* setelah proses pembersihan data ditunjukkan pada Gambar 3.2.

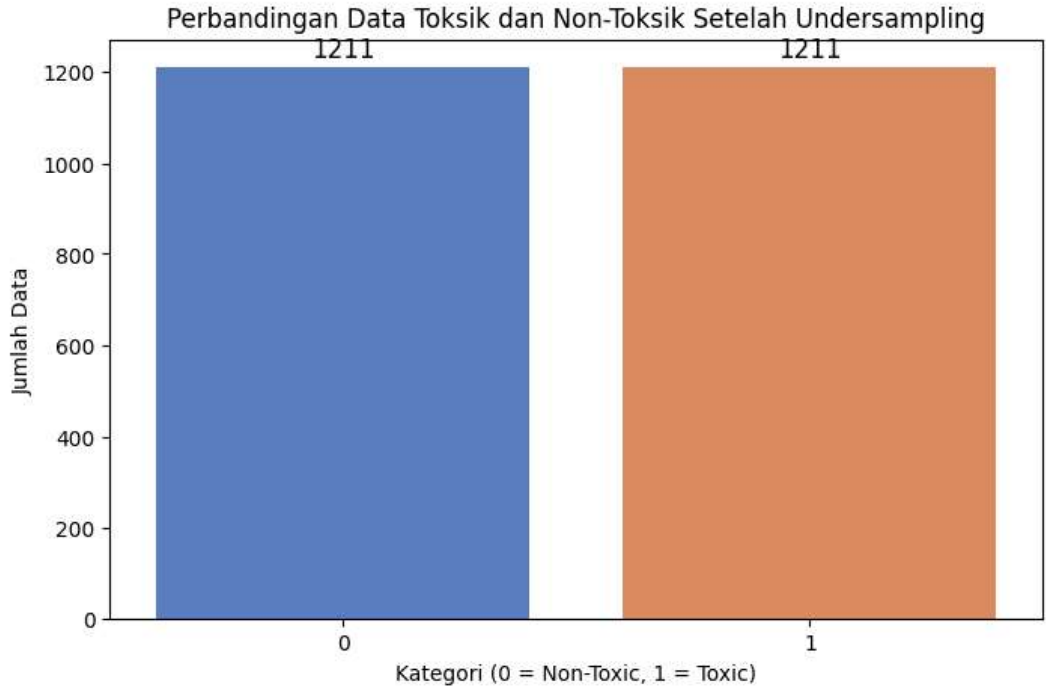


Gambar 4.1 Distribusi Label Level 1

Berdasarkan Gambar 3.2, dataset terdiri atas 18733 data *non-toxic* dan 1211 data *toxic*. Distribusi tersebut menunjukkan adanya ketidakseimbangan kelas yang cukup signifikan karena jumlah data *toxic* jauh lebih sedikit dibandingkan data *non-toxic*.

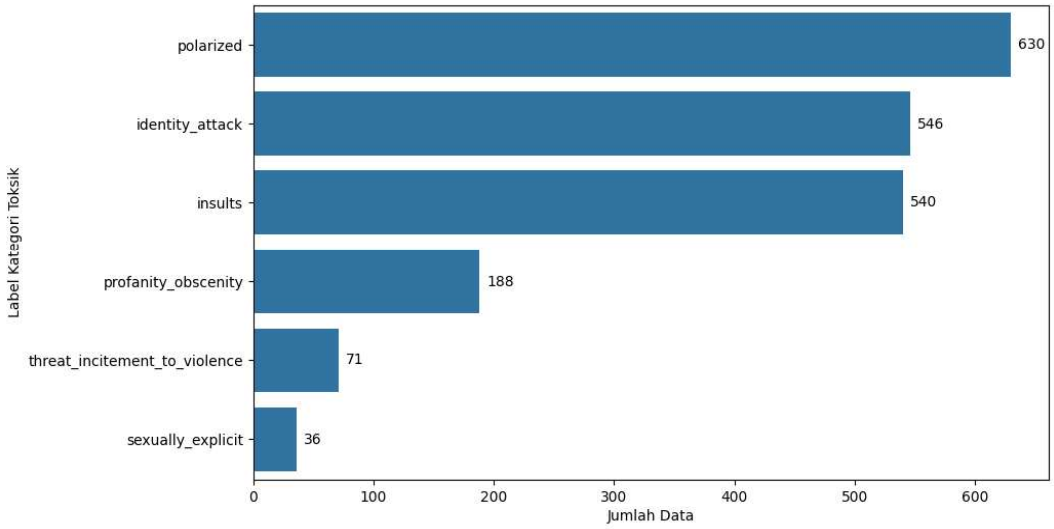
Untuk mengatasi ketidakseimbangan data pada label *toxicity*, penelitian ini menerapkan teknik *undersampling* pada kelas mayoritas sehingga diperoleh

distribusi kelas yang seimbang dengan rasio 1:1 antara kelas *toxic* dan *non-toxic*. Distribusi data setelah *undersampling* ditunjukkan pada Gambar 3.3.



Gambar 4.2 Distribusi Label Level 1 Setelah *Undersampling*

Berdasarkan Gambar 3.3, distribusi data setelah penerapan *undersampling* menjadi imbang dengan jumlah data per label 1211 data. Sementara itu, distribusi kategori ujaran toksik ditunjukkan pada Gambar 3.4.



Gambar 4.3 Distribusi kategori ujaran toksik

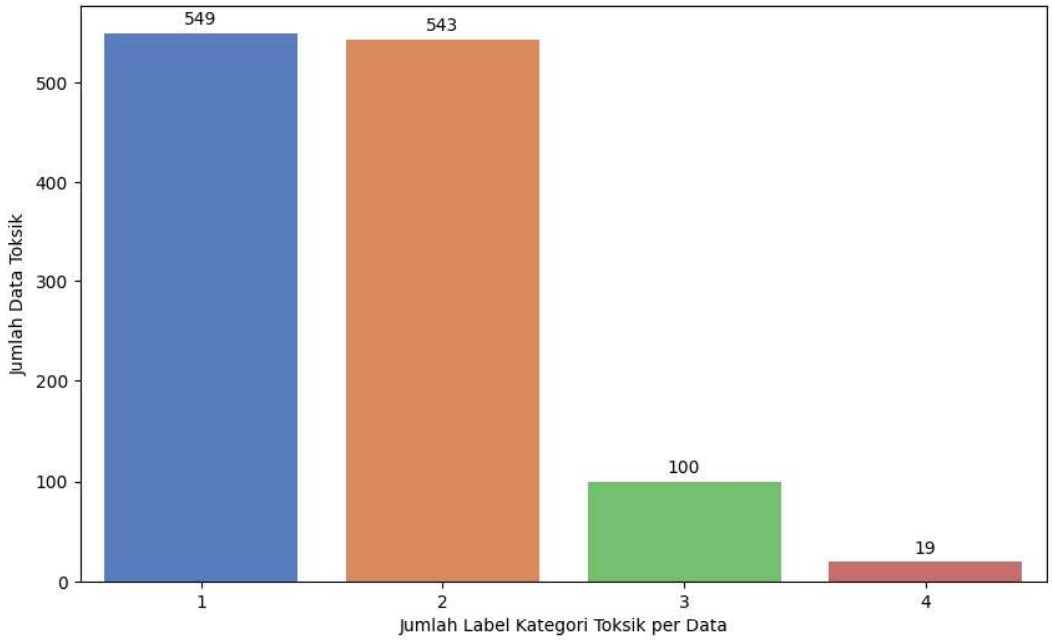
Berdasarkan Gambar 3.4, kategori *polarized* merupakan label yang paling banyak muncul dengan jumlah 630 data, diikuti oleh *identity attack* sebanyak 546 data dan *insults* sebanyak 540 data. Sebaliknya, kategori *sexually explicit* memiliki jumlah kemunculan paling sedikit, yaitu 36 data. Perbedaan jumlah kemunculan antar label menunjukkan bahwa ketidakseimbangan data tidak hanya terjadi pada label *toxicity*, tetapi juga pada masing-masing kategori ujaran toksik. Kondisi ini berpotensi menyebabkan model lebih cenderung mempelajari pola pada label yang memiliki jumlah sampel besar, sementara performa pada label dengan jumlah sampel sedikit dapat menurun karena kurangnya representasi data selama proses pelatihan.

Ketidakseimbangan data pada klasifikasi multilabel tidak dapat ditangani secara sederhana menggunakan teknik *undersampling* sebagaimana pada klasifikasi biner atau multikelas. Hal ini disebabkan satu sampel dapat memiliki lebih dari satu label secara bersamaan. Penghapusan sampel dari label mayoritas

berisiko menghilangkan informasi penting pada label lain yang mungkin termasuk kategori minoritas. Selain itu, proses *undersampling* dapat mengubah distribusi hubungan antar label yang merupakan karakteristik penting dalam tugas klasifikasi multilabel. Oleh karena itu, penerapan *undersampling* berpotensi mengurangi informasi yang tersedia dan menurunkan kemampuan model dalam mempelajari keterkaitan antar kategori toksik.

Untuk mengatasi permasalahan tersebut, penelitian ini menggunakan fungsi *Binary Cross Entropy with Logits Loss* (BCEWithLogitsLoss) pada proses pelatihan model. BCEWithLogitsLoss menghitung kesalahan prediksi secara independen untuk setiap label sehingga sesuai digunakan pada tugas klasifikasi multilabel. Selain itu, fungsi ini memungkinkan pemberian bobot yang berbeda pada masing-masing label melalui parameter *positive weight (pos_weight)*. Dengan mekanisme tersebut, kesalahan prediksi pada label yang memiliki jumlah sampel lebih sedikit akan diberikan penalti yang lebih besar dibandingkan label mayoritas. Pendekatan ini membantu model untuk lebih memperhatikan label-label minoritas, sehingga informasi yang terkandung dalam dataset tetap terjaga selama pelatihan.

Ketidakseimbangan tersebut menjadi semakin kompleks karena setiap sampel tidak hanya dapat memiliki satu kategori toksik, tetapi juga beberapa kategori secara bersamaan. Oleh karena itu, selain menganalisis distribusi masing-masing label, perlu dilakukan analisis terhadap jumlah kategori toksik yang muncul pada setiap sampel untuk memahami karakteristik multilabel pada dataset yang digunakan.



Gambar 4.4 Distribusi Jumlah Label Aktif Kategori Ujaran Toksik

Analisis terhadap jumlah label aktif pada setiap data toksik menunjukkan bahwa sebagian besar data memiliki satu hingga lebih kategori ujaran toksik secara bersamaan, sebagaimana ditunjukkan pada Gambar 3.5. Hal ini menunjukkan bahwa dataset memiliki karakteristik multilabel, di mana satu teks dapat mengandung lebih dari satu kategori ujaran toksik. Karakteristik tersebut menyebabkan hubungan antar label menjadi penting untuk dipertahankan selama proses pelatihan model. Oleh karena itu, penelitian ini tidak menggunakan teknik *undersampling* pada level kategori toksik karena berpotensi menghilangkan keterkaitan antar label dan mengubah distribusi kemunculan kombinasi label dalam dataset. Sebagai alternatif, penanganan ketidakseimbangan data dilakukan pada tahap pelatihan model menggunakan fungsi BCEWithLogitsLoss, yang dirancang

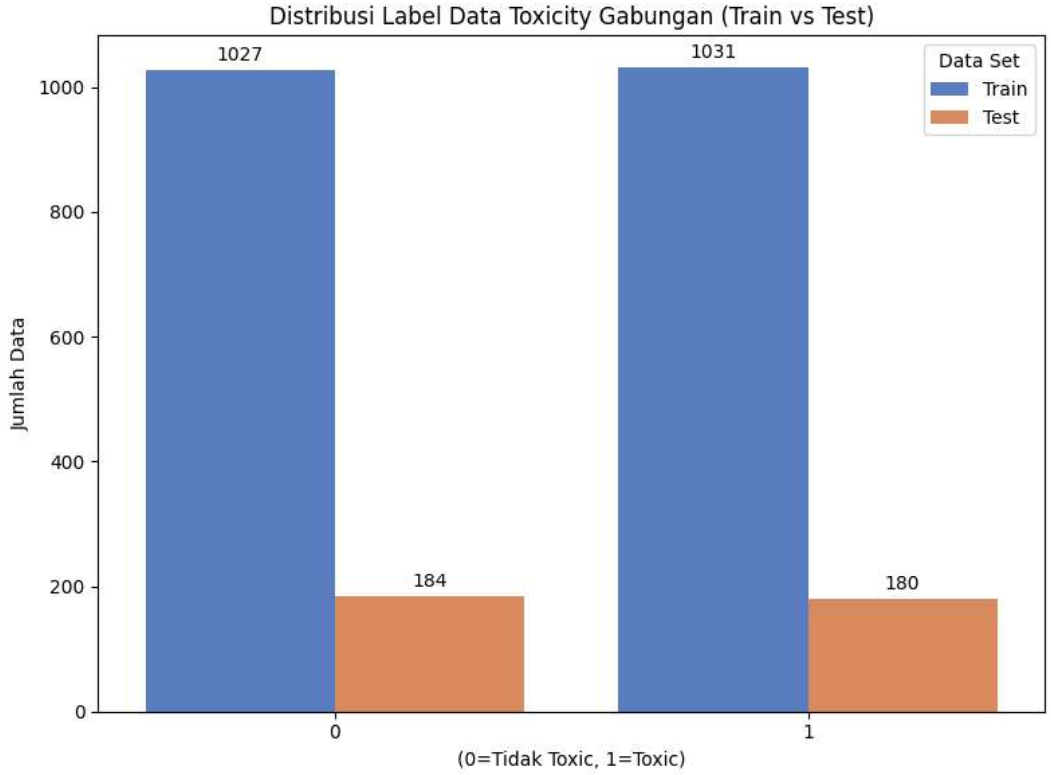
untuk klasifikasi multilabel dan mampu menghitung kesalahan prediksi secara independen pada setiap label tanpa mengubah distribusi data asli.

Dataset final dibagi menjadi 85% data pengembangan dan 15% hold-out test set. Data pengembangan digunakan untuk pelatihan dan validasi menggunakan Multilabel Stratified 5-Fold Cross Validation. Hold-out test set hanya digunakan sekali untuk evaluasi akhir dan perbandingan Skenario 2 dan Skenario 3. Pembagian ini menghasilkan 1.695 data pelatihan, 363 data validasi, dan 364 data pengujian. Pemisahan data dilakukan secara *stratified sampling* untuk mempertahankan proporsi kelas pada setiap subset sehingga distribusi label tetap merepresentasikan kondisi dataset secara keseluruhan. Data pengujian dipisahkan sejak awal dan tidak digunakan selama proses pelatihan maupun pemilihan model. Langkah ini bertujuan untuk menyediakan data yang benar-benar independen sehingga hasil evaluasi akhir dapat menggambarkan kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya.

Sementara itu, data pelatihan dan validasi digunakan pada tahap pengembangan model. Untuk memperoleh estimasi performa yang lebih stabil dan mengurangi pengaruh pembagian data tertentu, kedua subset tersebut kemudian digabungkan kembali dan dievaluasi menggunakan metode Stratified 5-Fold Cross Validation. Dengan pendekatan ini, setiap sampel pada data pelatihan dan validasi memperoleh kesempatan menjadi data validasi pada salah satu fold, sementara proporsi kelas tetap dipertahankan pada setiap pembagian fold.

Untuk memastikan bahwa proses pembagian data tidak menimbulkan bias distribusi, dilakukan analisis persebaran label pada masing-masing subset. Gambar

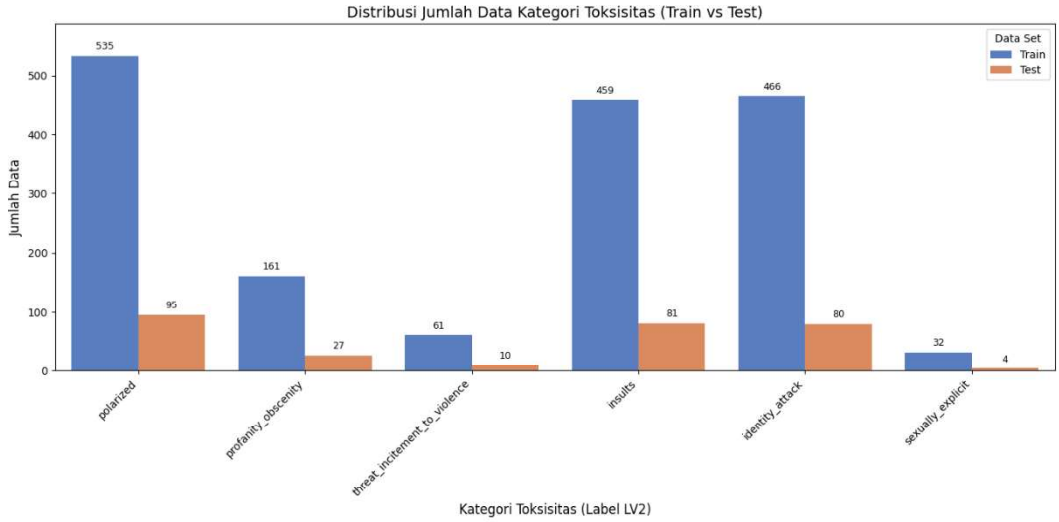
3.6 menunjukkan distribusi label *toxicity* pada data pelatihan, validasi, dan pengujian, sedangkan Gambar 3.7 menunjukkan distribusi masing-masing kategori toksisitas pada ketiga subset tersebut. Hasil analisis menunjukkan bahwa proporsi setiap label relatif konsisten antar subset, sehingga data pelatihan, validasi, dan pengujian dapat dianggap merepresentasikan distribusi dataset secara keseluruhan.



Gambar 4.5 Distribusi Label *toxicity* di Tiap Set

Berdasarkan Gambar 3.6, distribusi label *toxicity* pada data pelatihan dan pengujian menunjukkan proporsi yang relatif seimbang. Pada data pelatihan terdapat 1027 data berlabel tidak toksik dan 1031 data berlabel toksik. Sementara itu, data pengujian terdiri atas 184 data tidak toksik dan 180 data toksik. Hasil tersebut menunjukkan bahwa proses *stratified splitting* berhasil mempertahankan

proporsi kedua kelas pada setiap subset sehingga tidak terdapat perbedaan distribusi yang signifikan antara data pelatihan dan pengujian.



Gambar 4.6 Distribusi Label Kategori Toksisitas di Tiap Set

Selain distribusi label *toxicity*, pemeriksaan juga dilakukan terhadap distribusi masing-masing kategori toksisitas pada setiap subset data sebagaimana ditunjukkan pada Gambar 3.7. Kategori *polarized* memiliki jumlah kemunculan tertinggi pada seluruh subset, yaitu 535 data pada data pelatihan 95 data pada data pengujian. Kategori *insults* dan *identity attack* juga menunjukkan pola distribusi yang serupa dengan jumlah kemunculan yang relatif proporsional pada ketiga subset. Sementara itu, kategori *sexually explicit* tetap menjadi kategori dengan jumlah data paling sedikit, yaitu 32 data pada data pelatihan dan 4 data pada data pengujian.

Meskipun terdapat perbedaan jumlah kemunculan antar kategori toksisitas, pola distribusi setiap kategori pada data pelatihan, validasi, dan pengujian tetap menunjukkan proporsi yang konsisten. Sebagai contoh, kategori *profanity*

obscenity memiliki 161 data pada data pelatihan dan 27 data pada data pengujian, sedangkan kategori *threat incitement to violence* memiliki 61 data pada data pelatihan dan 10 data pada data pengujian. Kondisi ini menunjukkan bahwa proses pembagian data tidak menyebabkan suatu kategori terkonsentrasi pada subset tertentu. Dengan demikian, setiap subset tetap merepresentasikan karakteristik distribusi dataset secara keseluruhan.

Persebaran label yang relatif merata pada seluruh subset data menjadi penting untuk memastikan bahwa proses pelatihan, validasi, dan pengujian dilakukan pada data yang memiliki karakteristik serupa. Hal ini membantu mengurangi potensi bias evaluasi yang dapat muncul apabila terdapat perbedaan distribusi label yang signifikan antar subset. Selain itu, distribusi yang konsisten memungkinkan hasil pengukuran performa model mencerminkan kemampuan generalisasi model secara lebih akurat. Oleh karena itu, data pelatihan, validasi, dan pengujian yang dihasilkan dinilai telah memenuhi kebutuhan penelitian untuk digunakan pada tahap pengembangan dan evaluasi model klasifikasi toksisitas.

4.1.2 Definisi Tiga Skenario Pengujian

Penelitian ini mendefinisikan tiga skenario pengujian yang berbeda untuk mengevaluasi performa model XLM-RoBERTa dalam melakukan klasifikasi ujaran toksik pada komentar berbahasa Indonesia. Setiap skenario merepresentasikan pendekatan klasifikasi yang berbeda, yaitu *binary classification*, *flat multi-label classification*, dan *hierarchical multi-label classification*. Seluruh skenario menggunakan dataset yang sama, rasio pembagian data sebesar 85:15, metode penyeimbangan data menggunakan *Random Undersampling* dengan rasio

1:1, serta konfigurasi *hyperparameter* yang identik. Dengan demikian, perbedaan performa yang diperoleh dapat dikaitkan dengan perbedaan pendekatan klasifikasi yang digunakan.

a) *Binary Classification*

Skenario pertama menggunakan pendekatan *binary classification* untuk mengklasifikasikan komentar ke dalam dua kategori, yaitu *toxic* dan *non-toxic*. Pada skenario ini, seluruh kategori toksisitas digabungkan menjadi satu label biner sehingga model hanya bertugas mendeteksi keberadaan unsur toksik dalam suatu komentar tanpa mengidentifikasi jenis toksisitas yang terkandung di dalamnya. Skenario ini digunakan sebagai dasar pembanding (*baseline*) untuk mengukur kemampuan model dalam melakukan deteksi ujaran toksik secara umum.

b) *Flat Multi-Label Classification*

Skenario kedua menggunakan pendekatan *flat multi-label classification* yang memungkinkan model memprediksi seluruh kategori toksisitas secara langsung dalam satu proses klasifikasi. Setiap komentar dapat memiliki lebih dari satu label secara bersamaan sehingga model harus mempelajari hubungan antara teks komentar dan seluruh kategori toksisitas yang tersedia. Pendekatan ini menggunakan satu model XLM-RoBERTa untuk menghasilkan prediksi multilabel tanpa melalui tahapan klasifikasi sebelumnya.

c) *Hierarchical Multi-Label Classification*

Skenario ketiga menggunakan pendekatan *hierarchical multi-label classification* yang membagi proses klasifikasi ke dalam dua tingkat. Pada tingkat pertama, model XLM-RoBERTa digunakan untuk melakukan klasifikasi biner guna menentukan apakah suatu komentar termasuk *toxic* atau *non-toxic*. Selanjutnya, hanya komentar yang diprediksi sebagai *toxic* yang diteruskan ke tingkat kedua untuk dilakukan klasifikasi multilabel terhadap kategori-kategori toksisitas yang dimiliki. Dengan demikian, pendekatan ini menggunakan dua model XLM-RoBERTa yang berbeda dan disusun secara berurutan dalam sebuah *hierarchical pipeline*, di mana keluaran model pada tingkat pertama berfungsi sebagai penyaring sebelum proses identifikasi jenis toksisitas dilakukan pada tingkat kedua.

Ketiga skenario tersebut dievaluasi menggunakan metode *Stratified 5-Fold Cross Validation* pada data pelatihan dan validasi. Hasil evaluasi dari setiap *fold* kemudian dirata-ratakan untuk memperoleh nilai performa akhir yang lebih stabil dan representatif.

4.1.3 Metrik Evaluasi

Tahap evaluasi dilakukan untuk menilai performa model dalam melakukan klasifikasi hierarkis ujaran toksik menggunakan *FacebookAI/xlm-roberta-base*. Evaluasi ini bertujuan untuk mengukur sejauh mana sistem mampu mengenali dan mengklasifikasikan teks berbahasa Indonesia yang mengandung ujaran toksik secara tepat pada dua tingkat klasifikasi, yaitu klasifikasi toksik dan non-toksik pada level pertama serta klasifikasi multi-label jenis toksisitas pada level kedua.

Selain itu, evaluasi juga dilakukan secara *end-to-end* untuk mengukur kinerja sistem secara keseluruhan sebagai satu kesatuan proses.

a) Evaluasi Biner (Klasifikasi Toksik dan Non-Toksik)

Evaluasi biner bertujuan untuk mengukur kemampuan model dalam membedakan antara teks yang mengandung ujaran toksik dan teks yang bersifat non-toksik. Pada tahap ini, model menghasilkan nilai probabilitas toksisitas untuk setiap teks. Nilai probabilitas tersebut kemudian dibandingkan dengan nilai ambang batas sebesar 0,5 untuk menentukan prediksi akhir berupa label toksik atau non-toksik.

Hasil prediksi model selanjutnya dibandingkan dengan label sebenarnya untuk menghitung metrik evaluasi. Metrik yang digunakan pada level ini meliputi *accuracy*, *precision*, *recall*, dan *F1-score*. Keempat metrik tersebut digunakan secara bersamaan untuk memberikan gambaran yang seimbang mengenai performa model, khususnya dalam kondisi data yang tidak seimbang antara kelas toksik dan non-toksik.

Perhitungan metrik evaluasi biner dilakukan berdasarkan nilai *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN), sebagaimana dirumuskan pada Persamaan (3.13).

$$\begin{aligned}
\text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \\
\text{Precision} &= \frac{TP}{TP + FP} \\
\text{Recall} &= \frac{TP}{TP + FN} \\
\text{F1-Score} &= 2 \times \frac{Precision \times Recall}{Precision + Recall}
\end{aligned} \tag{3.13}$$

Keterangan:

TP (True Positive) = jumlah teks toksik yang diprediksi benar sebagai toksik,
TN (True Negative) = jumlah teks non-toksik yang diprediksi benar sebagai non-toksik,
FP (False Positive) = jumlah teks non-toksik yang salah diklasifikasikan sebagai toksik,
FN (False Negative) = jumlah teks toksik yang salah diklasifikasikan sebagai non-toksik.

Meskipun *accuracy* sering digunakan, metrik ini bisa menyesatkan jika digunakan pada dataset dengan distribusi kelas yang tidak seimbang, karena akurasi tinggi dapat diperoleh hanya dengan memprediksi kelas mayoritas. *Precision* menunjukkan proporsi prediksi positif untuk kelas tertentu yang benar-benar tepat. Nilai *Precision* yang tinggi berarti *model* jarang salah mengklasifikasikan kelas lain sebagai kelas tersebut. *Recall* mengukur kemampuan *model* untuk menangkap semua baris data dari kelas tertentu. Nilai tinggi berarti *model* berhasil mendeteksi sebagian besar contoh dari kelas tersebut. *F1-Score* adalah rata-rata harmonik dari *Precision* dan *Recall*, yang digunakan untuk menyeimbangkan keduanya dalam satu metrik. *F1-Score* sangat berguna ketika distribusi kelas tidak merata.

b) Evaluasi Multi-Label (Klasifikasi Multi-Label Jenis Toksisitas)

Evaluasi multi-label dilakukan untuk menilai kemampuan model dalam mengidentifikasi kategori toksisitas pada teks yang telah dinyatakan toksik pada level pertama. Oleh karena itu, evaluasi level kedua hanya diterapkan pada subset data uji yang memiliki label toksik. Pendekatan ini konsisten dengan desain klasifikasi hierarkis yang digunakan dalam penelitian ini.

Pada level ini, setiap teks dapat memiliki lebih dari satu label toksisitas secara bersamaan, sehingga digunakan fungsi aktivasi sigmoid untuk menghasilkan probabilitas independen bagi setiap kategori toksisitas. Kategori yang dievaluasi meliputi *polarized*, *profanity_obscenity*, *threat_incitement_to_violence*, *insults*,

identity_attack, dan *sexually_explicit*. Evaluasi dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score* untuk masing-masing label.

Selain evaluasi per label, penelitian ini juga menggunakan *Macro F1* untuk mengukur performa keseluruhan model pada klasifikasi multi-label. *Macro F1* diperoleh dengan menghitung nilai F1-score pada setiap label secara terpisah, kemudian mengambil rata-ratanya sebagaimana ditunjukkan pada Persamaan (3.14). Pendekatan ini memberikan bobot yang sama untuk setiap label sehingga tidak dipengaruhi oleh ketidakseimbangan distribusi data antar label.

$$Macro\ F1 = \frac{1}{L}\sum_{j=1}^{L} F\ 1_j \tag{3.14}$$

Keterangan:
L = jumlah label toksisitas yang dievaluasi.
$F1_j$ = nilai F1-score untuk label ke-j.

Sebagai contoh, apabila terdapat enam label toksisitas dengan nilai F1-score masing-masing sebesar 0,80; 0,75; 0,60; 0,70; 0,65; dan 0,50, maka nilai *Macro F1* dihitung sebagai:

$$Macro\ F1 = \frac{0{,}80 + 0{,}75 + 0{,}60 + 0{,}70 + 0{,}65 + 0{,}50}{6} = 0{,}67$$

Nilai tersebut menunjukkan rata-rata kemampuan model dalam mengenali seluruh kategori toksisitas secara seimbang. Selanjutnya ada *Subset Accuracy* digunakan untuk mengukur proporsi sampel yang seluruh labelnya berhasil diprediksi dengan benar oleh model. Pada klasifikasi multilabel, suatu prediksi dianggap benar hanya jika seluruh label hasil prediksi identik dengan seluruh label sebenarnya pada suatu sampel. Metrik ini merupakan salah satu metrik evaluasi

yang paling ketat karena kesalahan pada satu label saja akan menyebabkan prediksi untuk sampel tersebut dianggap salah.

$$\text{SA} = \frac{1}{N}\sum_{i=1}^{N} I(y_i = \hat{y}_i) \tag{3.15}$$

Keterangan:
N = jumlah sampel.
i = indeks sampel ke-i.
Y_i = himpunan label sebenarnya (*ground truth*) pada sampel ke-i.
$\hat{Y}_i$ = himpunan label hasil prediksi model pada sampel ke-i.
$I(Y_i = \hat{Y}_i)$ = fungsi indikator yang bernilai:
1, jika seluruh label prediksi sama persis dengan seluruh label sebenarnya pada sampel ke-i;
0, jika terdapat minimal satu label yang berbeda.

Pada klasifikasi multilabel, satu sampel dapat memiliki lebih dari satu label secara bersamaan. Oleh karena itu, evaluasi tidak hanya mempertimbangkan kebenaran setiap label secara individu, tetapi juga kesesuaian seluruh kombinasi label yang diprediksi. *Subset Accuracy* digunakan untuk mengukur kemampuan model dalam memprediksi keseluruhan himpunan label secara tepat pada setiap sampel. Nilainya berada pada rentang 0 hingga 1. Semakin tinggi nilainya, semakin baik performa model dalam menghasilkan prediksi yang sepenuhnya sesuai dengan label sebenarnya. Sebaliknya, nilai *Subset Accuracy* yang rendah menunjukkan bahwa model masih sering melakukan kesalahan pada satu atau lebih label dalam suatu sampel.

Yang terakhir, metrik *Hamming Loss* digunakan untuk mengukur rata-rata kesalahan prediksi pada setiap label dalam tugas klasifikasi multilabel. Berbeda dengan *Subset Accuracy* yang hanya memberikan nilai benar apabila seluruh label pada suatu sampel diprediksi secara tepat, *Hamming Loss* mengevaluasi kesesuaian prediksi pada setiap label secara individual. Dengan demikian, metrik ini dapat

memberikan informasi yang lebih rinci mengenai tingkat kesalahan model pada masing-masing label. Nilai *Hamming Loss* berada pada rentang 0 hingga 1, di mana nilai yang semakin mendekati 0 menunjukkan semakin sedikit kesalahan prediksi yang dilakukan model, sedangkan nilai yang mendekati 1 menunjukkan tingkat kesalahan yang semakin tinggi.

Hamming Loss dihitung berdasarkan proporsi label yang salah diprediksi terhadap seluruh kombinasi sampel dan label yang dievaluasi. Pada setiap pasangan label aktual dan label prediksi dilakukan operasi XOR (Exclusive OR). Operasi tersebut menghasilkan nilai 1 apabila prediksi tidak sesuai dengan label sebenarnya dan nilai 0 apabila prediksi sesuai. Seluruh nilai kesalahan kemudian dijumlahkan dan dibagi dengan total kombinasi sampel dan label sehingga diperoleh rata-rata kesalahan per label. Perhitungan *Hamming Loss* ditunjukkan pada Persamaan (3.16).

$$\text{Hamming Loss} = \frac{1}{N \times L} \sum_{i=1}^{N} \sum_{j=1}^{L} XOR(y_{ij}, \hat{y}_{ij}) \tag{3.16}$$

Keterangan:
N = jumlah sampel,
L = jumlah label,
y_{ij} = label sebenarnya untuk sampel ke-i dan label ke-j,
$\hat{y}_{ij}$ = hasil prediksi model untuk sampel ke-i dan label ke-j,
XOR = operasi logika yang menghasilkan 1 jika prediksi salah dan 0 jika benar.

Sebagai contoh, misalkan terdapat 100 teks toksik pada data uji dengan 6 label toksisitas. Total kombinasi evaluasi adalah 600 label. Jika model menghasilkan 90 kesalahan prediksi label, maka nilai *Hamming Loss* adalah 90 / 600 = 0,15. Nilai ini menunjukkan bahwa rata-rata kesalahan prediksi per label

relatif rendah, sehingga performa model pada klasifikasi multi-label dapat dikatakan baik.

c) Evaluasi End-to-End (E2E)

Evaluasi *end-to-end* dilakukan untuk mengukur kinerja sistem klasifikasi secara keseluruhan, dengan mempertimbangkan seluruh alur hierarkis mulai dari level pertama hingga level kedua. Evaluasi ini penting karena kesalahan pada level pertama secara langsung akan memengaruhi hasil klasifikasi pada level kedua. Pada evaluasi *end-to-end*, sebuah prediksi dianggap benar apabila model berhasil memprediksi label toksik atau non-toksik dengan benar pada level pertama, dan apabila teks tersebut bersifat toksik, model juga harus berhasil memprediksi seluruh label jenis toksisitas yang sesuai pada level kedua. Jika terjadi kesalahan pada salah satu level, maka prediksi tersebut dianggap salah secara *end-to-end*.

Tabel 4.2 Contoh Data Evaluasi E2E

Sampel	**Level 1 Toxicity**		**Level 2 Kategori Toksisitas**		**Hasil**
	Aktual	**Prediksi**	**Aktual**	**Prediksi**	
1	1	1	[1, 0, 1, 0, 0]	[1, 0, 1, 0, 0]	Benar (1)
2	0	0	[0, 0, 0, 0, 0]	[0, 0, 0, 0, 0]	Benar (1)
3	1	1	[1, 1, 0, 0, 1]	[1, 1, 0, 1, 1]	Salah (0)
4	1	0	[1, 0, 0, 1, 0]	[0, 0, 0, 0, 0]	Salah (0)

Berdasarkan Tabel 4.2, sampel pertama memperoleh nilai 1 karena prediksi pada Level 1 dan seluruh label pada Level 2 sesuai dengan label aktual. Sampel kedua juga memperoleh nilai 1 karena model berhasil mengklasifikasikan teks sebagai non-toksik pada Level 1. Sampel ketiga memperoleh nilai 0 karena terdapat perbedaan pada salah satu label toksisitas meskipun hasil klasifikasi Level 1 sudah benar. Sampel keempat memperoleh nilai 0 karena model gagal mengidentifikasi

teks sebagai toksik pada Level 1 sehingga proses klasifikasi jenis toksisitas tidak dapat dilakukan dengan benar.

Hal ini menunjukkan bahwa metrik E2E Accuracy hanya memberikan nilai benar apabila seluruh tahapan klasifikasi hierarkis berhasil diprediksi secara tepat, mulai dari klasifikasi toxic/non-toxic pada Level 1 hingga identifikasi seluruh jenis toksisitas pada Level 2. Nilai E2E Accuracy dihitung sebagai:

$$E2E\ Accuracy = \frac{2}{4} = 0{,}5$$

Nilai tersebut menunjukkan bahwa sistem berhasil menyelesaikan keseluruhan proses klasifikasi hierarkis dengan benar pada 2 dari 4 sampel yang dievaluasi.

4.2 Hasil Uji Coba

Penelitian ini melakukan pengujian pada tiga skenario yang berbeda. Skenario pertama fokus pada klasifikasi biner, yaitu menentukan apakah suatu teks termasuk *toxic* atau *non-toxic*. Skenario kedua menggunakan pendekatan *flat multilabel*, di mana model memprediksi seluruh label toksisitas secara bersamaan tanpa struktur hierarki. Skenario ketiga menerapkan pendekatan *hierarchical multilabel classification*, di mana Level 1 bertugas menentukan apakah teks mengandung toksisitas atau tidak, dan Level 2 mengidentifikasi jenis toksisitas apabila teks terdeteksi mengandung unsur *toxic*.

Seluruh skenario menggunakan konfigurasi eksperimen yang seragam untuk memastikan hasil pengujian dapat dibandingkan secara objektif. Model dilatih dengan *random seed* sebesar 42, dengan jumlah *epoch* untuk Level 1

sebanyak 10, Level 2 sebanyak 20, sedangkan untuk skenario biner dan *flat multilabel* masing-masing dilakukan selama 20 *epoch*. Strategi *early stopping* diterapkan dengan nilai tiga *epoch* untuk mencegah *overfitting*, sedangkan *batch size* yang digunakan adalah 16 untuk pelatihan dan 64 untuk evaluasi. Pembelajaran dilakukan dengan *learning rate* 2e-5 untuk skenario 1, skenario 2, dan skenario 3 Level 1 dan 1e-5 untuk skenario 3 Level 2. Panjang maksimal teks (*MAX_LENGTH*) mengikuti penelitian sebelumnya yakni 512. Data dibagi dengan proporsi 70% untuk pelatihan, 15% untuk validasi, dan 15% untuk pengujian, dan strategi *undersampling* diterapkan agar distribusi kelas seimbang dengan rasio 1:1. Model yang digunakan adalah *FacebookAI/xlm-roberta-base*, yang telah terlatih pada berbagai bahasa sehingga mampu menangani teks berbahasa Indonesia maupun campuran bahasa dengan efektif.

4.2.1 Hasil Uji Skenario Klasifikasi Biner

Skenario Biner bertujuan untuk mengevaluasi kemampuan model RoBERTa mengidentifikasi apakah suatu teks termasuk kategori *toxic* atau *non-toxic*. Pengujian dilakukan menggunakan metode *5-Fold Stratified Cross Validation* sebagaimana diterapkan pada penelitian rujukan. Pada setiap fold, model dilatih menggunakan empat bagian data dan diuji menggunakan satu bagian data yang berbeda, sehingga seluruh data memperoleh kesempatan untuk digunakan sebagai data pengujian. Hasil pengujian pada masing-masing fold ditunjukkan pada Tabel 4.3.

Tabel 4.3 Hasil Evaluasi Skenario Klasifikasi Biner

Fold	**Accuracy**	**Precision**	**Recall**	**F1-Score**
1	80,94%	78,28%	85,66%	81,80%
2	79,25%	77,39%	82,64%	79,93%
3	81,32%	79,43%	84,53%	81,90%
4	79,40%	75,49%	87,17%	80,91%
5	75,05%	69,88%	87,88%	77,85%
Rata-rata	79,19%	76,09%	85,58%	80,48%

Berdasarkan Tabel 4.3, model memperoleh rata-rata *accuracy* sebesar 79,19%, *precision* sebesar 76,09%, *recall* sebesar 85,58%, dan *F1-Score* sebesar 80,48%. Nilai *accuracy* menunjukkan bahwa sebagian besar data berhasil diklasifikasikan dengan benar oleh model. Sementara itu, nilai *F1-Score* yang berada di atas 80% menunjukkan bahwa model mampu menjaga keseimbangan antara kemampuan mendeteksi data *toxic* dan meminimalkan kesalahan klasifikasi.

Nilai *recall* yang lebih tinggi dibandingkan *precision* menunjukkan bahwa model memiliki kemampuan yang baik dalam mengenali data *toxic*. Dengan rata-rata *recall* sebesar 85,58%, sebagian besar teks yang mengandung ujaran toksik berhasil terdeteksi oleh model. Namun demikian, nilai *precision* yang sedikit lebih rendah mengindikasikan bahwa masih terdapat sejumlah teks yang diprediksi sebagai *toxic* meskipun sebenarnya termasuk kategori *non-toxic*. Kondisi ini menunjukkan adanya *trade-off* antara kemampuan mendeteksi sebanyak mungkin ujaran toksik dan risiko menghasilkan prediksi positif yang kurang tepat.

Jika ditinjau dari hasil pada masing-masing fold, nilai *F1-Score* berada pada rentang 77,85% hingga 81,90%. Perbedaan nilai yang relatif kecil antar fold menunjukkan bahwa model memiliki performa yang cukup stabil terhadap variasi pembagian data yang dihasilkan oleh proses *cross validation*. Selain itu, seluruh

fold menghasilkan nilai *recall* di atas 82%, yang menunjukkan konsistensi model dalam mendeteksi teks yang mengandung ujaran toksik.

Secara keseluruhan, hasil pengujian pada Skenario 1 menunjukkan bahwa model RoBERTa mampu melakukan klasifikasi *toxic* dan *non-toxic* dengan baik. Nilai *F1-Score* sebesar 80,48% serta performa yang relatif stabil pada seluruh fold menunjukkan bahwa model memiliki kemampuan yang memadai dalam mendeteksi keberadaan ujaran toksik pada teks berbahasa Indonesia. Hasil ini menjadi acuan untuk membandingkan performa model pada skenario pengujian lainnya.

4.2.2 Hasil Uji Skenario Flat Multilabel Classification

Skenario *Flat Multilabel Classification* bertujuan untuk mengevaluasi performa model menggunakan pendekatan *flat multi-label classification*. Pada pendekatan ini, model secara langsung memprediksi seluruh label yang tersedia dalam satu tahap klasifikasi tanpa melalui proses klasifikasi bertingkat. Dengan demikian, model harus mempelajari hubungan antara teks dan seluruh label toksisitas secara bersamaan. Pengujian dilakukan menggunakan metode *5-Fold Stratified Cross Validation*. Pada setiap fold, model dilatih menggunakan empat bagian data dan diuji menggunakan satu bagian data yang berbeda. Hasil evaluasi pada masing-masing fold ditunjukkan pada Tabel 4.4.

Tabel 4.4 Hasil Evaluasi Skenario *Flat Multilabel Classification*

Fold	**Accuracy**	**F1-Macro**	**Hamming Loss**
1	39,32%	34,14%	15,05%
2	44,17%	34,11%	13,22%
3	46,60%	32,02%	20,67%
4	39,66%	25,04%	15,05%
5	41,85%	35,04%	13,84%
Rata-rata	42,32% ± 4%	32,07% ± 3%	14,02% ± 5%

Berdasarkan Tabel 4.4, model memperoleh rata-rata *accuracy* sebesar 42,32% dan *F1-Macro* sebesar 32,07%. Nilai *F1-Macro* digunakan sebagai metrik utama karena permasalahan yang dihadapi merupakan klasifikasi multi-label dengan distribusi data yang tidak seimbang pada setiap kategori. Nilai tersebut menunjukkan bahwa model masih mengalami kesulitan dalam mengenali seluruh label toksisitas secara merata. Berdasarkan hasil *5-fold cross-validation*, model menghasilkan nilai *F1-Macro* sebesar 32,07% ± 4,08% dan *accuracy* sebesar 42,32% ± 3,08%. Nilai standar deviasi yang relatif rendah pada kedua metrik tersebut menunjukkan bahwa performa model cenderung konsisten pada setiap *fold* pengujian. Hal ini mengindikasikan bahwa model memiliki tingkat stabilitas yang cukup baik terhadap variasi pembagian data, sehingga hasil evaluasi yang diperoleh dapat dianggap representatif terhadap kemampuan model secara keseluruhan.

Rendahnya nilai *F1-Macro* dibandingkan dengan skenario klasifikasi biner menunjukkan bahwa tugas klasifikasi multi-label memiliki tingkat kompleksitas yang lebih tinggi. Pada pendekatan *flat multi-label classification*, model harus secara langsung menentukan keberadaan beberapa kategori toksisitas dalam satu proses prediksi. Selain itu, perbedaan jumlah data pada masing-masing label juga dapat memengaruhi kemampuan model dalam mempelajari karakteristik setiap kategori secara optimal.

Secara keseluruhan, hasil pengujian pada Skenario 2 menunjukkan bahwa pendekatan *flat multi-label classification* mampu melakukan klasifikasi terhadap beberapa kategori ujaran toksik secara bersamaan, namun performa yang diperoleh masih relatif rendah. Hasil ini akan dibandingkan dengan pendekatan klasifikasi hierarkis pada skenario berikutnya untuk mengetahui pendekatan yang memberikan performa lebih baik dalam tugas klasifikasi ujaran toksik multi-label.

4.2.3 Hasil Uji Skenario Hierarchical Multilabel Classification

Skenario *hierarchical classification* membagi proses klasifikasi ke dalam dua tahap. Pada tahap pertama (Level 1), model melakukan klasifikasi biner untuk menentukan apakah suatu teks termasuk kategori *toxic* atau *non-toxic*. Selanjutnya, teks yang diprediksi sebagai *toxic* akan diteruskan ke tahap kedua (Level 2) untuk dilakukan klasifikasi multi-label guna mengidentifikasi jenis toksisitas yang terkandung di dalamnya.

Pengujian pada skenario ini dilakukan menggunakan metode *5-Fold Stratified Cross Validation*. Berbeda dengan Skenario flat yang melakukan prediksi seluruh label secara langsung dalam satu model, pendekatan hierarkis memisahkan proses klasifikasi label *toxicity* dan klasifikasi kategori toksik ke dalam dua model yang berbeda. Dengan demikian, setiap model dapat berfokus pada tugas klasifikasi yang lebih spesifik.

Hasil Evaluasi Level 1

Level 1 bertugas mengidentifikasi keberadaan unsur toksik pada suatu teks. Hasil evaluasi menggunakan *5-Fold Stratified Cross Validation* ditunjukkan pada Tabel 4.5.

Tabel 4.5 Hasil Evaluasi Level 1 pada Skenario 3

Fold	**Accuracy**	**Precision**	**Recall**	**F1-Score**
1	79,61%	76,62%	85,51%	80,82%
2	79,13%	80,30%	77,18%	78,71%
3	81,80%	81,04%	83,01%	82,01%
4	77,13%	77,18%	77,18%	77,18%
5	80,54%	78,38%	84,47%	81,31%
Rata-rata	79,64% ± 1%	78,71% ± 1%	81,47% ± 4%	80,01% ± 2%

Berdasarkan Tabel 4.5, model memperoleh rata-rata *accuracy* sebesar 79,64%, *precision* sebesar 78,71%, *recall* sebesar 81,47%, dan *F1-Score* sebesar 80,01%. Hasil tersebut menunjukkan bahwa model mampu membedakan teks *toxic* dan *non-toxic* dengan cukup baik. Nilai *recall* yang lebih tinggi dibandingkan *precision* mengindikasikan bahwa model cenderung lebih sensitif dalam mendeteksi teks toksik.

Karakteristik ini sesuai dengan tujuan pendekatan hierarkis. Pada sistem hierarkis, kesalahan pada Level 1 akan berdampak langsung terhadap proses selanjutnya karena teks yang diprediksi sebagai *non-toxic* tidak akan diproses oleh Level 2. Oleh karena itu, kemampuan model untuk mengenali sebanyak mungkin teks toksik menjadi aspek yang penting. Nilai *recall* sebesar 81,47% menunjukkan bahwa sebagian besar teks toksik berhasil terdeteksi dan dapat diteruskan ke tahap klasifikasi jenis toksisitas.

Selain itu, performa model pada setiap fold menunjukkan variasi yang relatif kecil. Nilai standar deviasi *F1-Score* yang relatif rendah di ± 2% yang mengindikasikan bahwa model memiliki tingkat konsistensi yang cukup baik terhadap variasi pembagian data. Hal ini menunjukkan bahwa model mampu mempelajari pola umum ujaran toksik secara stabil dan tidak terlalu bergantung pada karakteristik data pada fold tertentu.

Hasil Evaluasi Level 2

Setelah suatu teks diketahui bersifat toxic, proses klasifikasi dilanjutkan ke Level 2 untuk mengidentifikasi jenis toksisitas yang terkandung di dalam teks tersebut. Berbeda dengan Level 1, proses pelatihan dan evaluasi pada Level 2 hanya menggunakan data yang memiliki label *toxicity* = 1. Dengan demikian, hasil evaluasi pada tahap ini menggambarkan kemampuan model dalam mengidentifikasi kategori toksisitas ketika keberadaan unsur toksik telah diketahui sebelumnya. Mengingat tugas pada Level 2 merupakan klasifikasi *multi-label*, proses *cross-validation* dilakukan menggunakan *Multilabel Stratified K-Fold*. Metode ini dipilih untuk menjaga distribusi masing-masing label toksisitas tetap seimbang pada setiap *fold*, sehingga proporsi kemunculan setiap kategori toksisitas pada data pelatihan dan data pengujian relatif serupa. Dengan distribusi data yang lebih representatif, hasil evaluasi yang diperoleh dapat memberikan gambaran yang lebih stabil dan akurat mengenai kemampuan model dalam mengidentifikasi berbagai jenis toksisitas. Hasil evaluasi pada Level 2 ditunjukkan pada Tabel 4.6.

Tabel 4.6 Hasil Evaluasi Level 2 pada Skenario 3

Fold	**Subset Accuracy**	**F1-Macro**	**Hamming Loss**
1	4,98%	55,90%	28,69%
2	2,80%	46,20%	32,32%
3	3,92%	50,77%	31,86%
4	4,37%	54,61%	32,36%
5	4,37%	54,86%	33,74%
Rata-rata	4% ± 0,8%	52,47% ± 4%	31,79% ± 1,8%
Label	**Precision**	**Recall**	**F1-Score**
polarized	55,22%	97,94%	70,57%
profanity_obscenity	33,65%	59,58%	42,46%
Threat_incitement _to_violence	29,08%	16,28%	13,99%
insults	44,54%	100,00%	61,62%
Identity_attack	45,26%	99,79%	62,27%
sexually_explicit	65,95%	75,71%	63,90%

Nilai *F1-Macro* digunakan sebagai metrik utama karena memberikan bobot yang sama kepada setiap label tanpa dipengaruhi jumlah sampel pada masing-masing kategori. Hasil *F1-Macro* sebesar 52,47% menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam mengenali berbagai jenis toksisitas yang terdapat pada data berlabel *toxic*.

Berdasarkan Tabel 4.6, performa terbaik diperoleh pada label *polarized* dengan nilai *F1-Score* sebesar 70,57%. Tingginya nilai *recall* sebesar 97,94% menunjukkan bahwa sebagian besar teks yang mengandung unsur polarisasi berhasil dikenali oleh model. Hasil ini mengindikasikan bahwa pola bahasa pada kategori *polarized* relatif lebih konsisten sehingga lebih mudah dipelajari dibandingkan kategori lainnya.

Label *identity attack* memperoleh *F1-Score* sebesar 62,27%, sedangkan label *insults* memperoleh *F1-Score* sebesar 61,62%. Kedua label tersebut juga memiliki nilai *recall* yang sangat tinggi, masing-masing sebesar 99,79% dan 100%. Hasil ini menunjukkan bahwa model sangat jarang gagal mendeteksi keberadaan

kedua kategori tersebut. Namun demikian, nilai *precision* yang masih relatif rendah menunjukkan bahwa model masih menghasilkan sejumlah prediksi positif yang sebenarnya bukan termasuk kategori tersebut.

Pada label *sexually explicit*, model memperoleh *F1-Score* sebesar 63,90%. Nilai tersebut merupakan yang tertinggi kedua setelah *polarized*. Meskipun jumlah sampel pada kategori ini relatif lebih sedikit dibandingkan beberapa label lainnya, model masih mampu mempertahankan performa yang baik. Hal ini menunjukkan bahwa karakteristik bahasa pada kategori tersebut memiliki pola yang lebih spesifik sehingga lebih mudah dikenali oleh model.

Sementara itu, label *profanity/obscenity* memperoleh *F1-Score* sebesar 42,46%. Nilai ini menunjukkan bahwa model masih mengalami kesulitan dalam membedakan penggunaan kata-kata kasar yang benar-benar bersifat toksik dengan kata-kata yang muncul dalam konteks lain. Variasi penggunaan bahasa pada kategori ini menyebabkan proses klasifikasi menjadi lebih menantang.

Performa terendah diperoleh pada label *threat/incitement to violence* dengan nilai *F1-Score* sebesar 13,99%. Rendahnya performa pada kategori ini kemungkinan dipengaruhi oleh jumlah data yang sangat sedikit dibandingkan label lainnya. Berdasarkan distribusi data pada Bab III, label *threat/incitement to violence* hanya memiliki 82 sampel sehingga model memiliki keterbatasan dalam mempelajari karakteristik bahasa yang merepresentasikan ancaman atau ajakan kekerasan.

Secara keseluruhan, hasil *multilabel stratified 5-fold cross-validation* pada Level 2 menghasilkan nilai *F1-Macro* sebesar 52,47% ± 4,04%. Nilai ini

menunjukkan bahwa ketika seluruh label diberi bobot yang sama, model mampu mencapai performa yang cukup baik dalam mengidentifikasi berbagai jenis toksisitas. Nilai standar deviasi yang relatif rendah juga menunjukkan bahwa performa model cenderung konsisten pada setiap *fold* pengujian.

Selain itu, model memperoleh nilai *Hamming Loss* sebesar 31,79% ± 1,88%. Nilai ini menunjukkan proporsi kesalahan prediksi label terhadap seluruh label yang dievaluasi. Semakin kecil nilai *Hamming Loss*, semakin sedikit kesalahan klasifikasi yang dilakukan model pada setiap label. Hasil yang diperoleh menunjukkan bahwa meskipun model masih melakukan kesalahan pada beberapa kategori, performanya relatif stabil pada seluruh *fold*.

Di sisi lain, nilai *Subset Accuracy* yang diperoleh hanya sebesar 4,09% ± 0,84%. Nilai ini menunjukkan bahwa hanya sebagian kecil sampel yang seluruh labelnya dapat diprediksi dengan tepat secara bersamaan. Rendahnya nilai *Subset Accuracy* merupakan kondisi yang umum terjadi pada permasalahan klasifikasi *multi-label*, terutama ketika jumlah kombinasi label yang mungkin sangat beragam dan distribusi data antar label tidak seimbang.

Secara keseluruhan, hasil pengujian menunjukkan bahwa pendekatan hierarkis mampu mempertahankan performa yang baik pada tahap deteksi toksisitas dengan *F1-Score* sebesar 80,01% pada Level 1. Sementara itu, pada Level 2 model mampu mengidentifikasi beberapa kategori toksisitas dengan cukup baik, terutama pada kategori *polarized*, *sexually explicit*, *identity attack*, dan *insults*. Perlu diperhatikan bahwa hasil pada Level 2 merupakan evaluasi terhadap model klasifikasi jenis toksisitas yang dilakukan menggunakan data *toxic* saja, sehingga

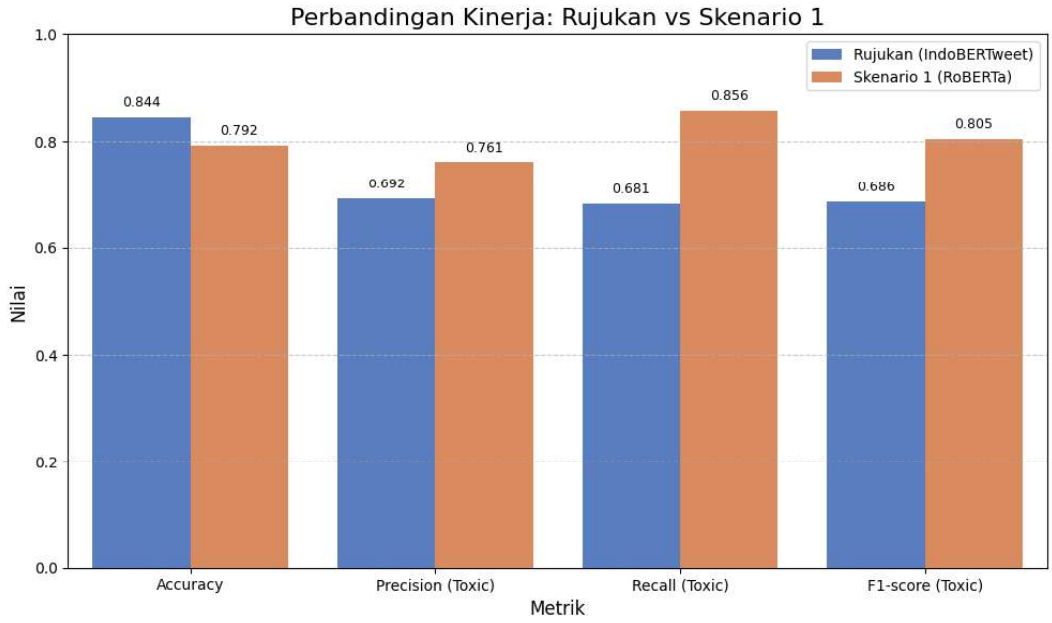
hasil tersebut belum merepresentasikan performa sistem hierarkis secara keseluruhan. Evaluasi *end-to-end* terhadap keseluruhan alur klasifikasi hierarkis akan dibahas pada pengujian menggunakan *test set* pada subbab berikutnya.

4.3 Perbandingan Hasil Test Antar Skenario

Setelah seluruh skenario pengujian dilakukan, tahap berikutnya adalah membandingkan hasil yang diperoleh dari masing-masing pendekatan. Perbandingan ini penting karena setiap skenario memiliki tujuan dan karakteristik yang berbeda. Skenario 1 hanya berfokus pada klasifikasi biner untuk membedakan teks *toxic* dan *non-toxic*, sedangkan Skenario 2 dan Skenario 3 sudah masuk ke permasalahan yang lebih kompleks, yaitu klasifikasi multilabel. Oleh karena itu, perbandingan tidak hanya dilihat dari angka metrik secara langsung, tetapi juga dari kesesuaian pendekatan dengan tujuan penelitian, yaitu mendeteksi ujaran toksik sekaligus mengenali jenis toksisitas yang muncul pada suatu teks.

4.3.1 Perbandingan Skenario Klasifikasi Biner dengan Model Rujukan

Perbandingan pada subbab ini dilakukan antara Skenario 1 menggunakan model *RoBERTa* dengan model rujukan yaitu *IndoBERTweet*. Model rujukan dipilih karena telah digunakan secara luas sebagai baseline dalam tugas klasifikasi teks berbahasa Indonesia, khususnya pada domain deteksi ujaran toksik. Evaluasi dilakukan berdasarkan metrik hasil rata-rata 5-*fold cross validation* pada Skenario 1 dan hasil evaluasi yang dilaporkan pada penelitian rujukan. Metrik yang digunakan meliputi *accuracy*, *precision*, *recall*, *F1-score*, *F1-macro*.



Gambar 4.7 Perbandingan Kinerja: Rujukan vs Skenario 1

Hasil perbandingan pada Gambar 4.1 menunjukkan bahwa *IndoBERTweet* memiliki nilai *accuracy* yang lebih tinggi dibandingkan Skenario 1. Nilai *accuracy* pada *IndoBERTweet* mencapai 84,4%, sedangkan Skenario 1 memperoleh 79,19%. Hal ini menunjukkan bahwa model rujukan lebih baik dalam menghasilkan prediksi yang benar secara keseluruhan pada data uji. Namun demikian, selisih nilai tersebut tidak terlalu besar sehingga menunjukkan bahwa performa kedua model masih berada pada tingkat yang relatif berdekatan.

Pada metrik kelas toxic, Skenario 1 menunjukkan performa yang lebih unggul pada seluruh indikator utama. Nilai *precision* Skenario 1 sebesar 76,09% lebih tinggi dibandingkan *IndoBERTweet* yang sebesar 69,2%, yang menunjukkan bahwa Skenario 1 lebih sedikit menghasilkan *false positive*. Selain itu, nilai *recall* Skenario 1 mencapai 85,58% yang jauh lebih tinggi dibandingkan 68,1% pada model rujukan. Hal ini menunjukkan bahwa Skenario 1 lebih sensitif dalam

mendeteksi teks yang mengandung ujaran toksik sehingga lebih sedikit kasus toxic yang tidak terdeteksi.

Pada metrik *F1-score* untuk kelas toxic, Skenario 1 memperoleh nilai 80,48% sedangkan *IndoBERTweet* memperoleh 68,6%. Perbedaan ini menunjukkan bahwa Skenario 1 memiliki keseimbangan yang lebih baik antara *precision* dan *recall* pada kelas toxic. Namun demikian, pada metrik *F1-macro*, kedua model menunjukkan nilai yang hampir sama, yaitu 79,07% pada Skenario 1 dan 79,1% pada *IndoBERTweet*. Hal ini mengindikasikan bahwa secara keseluruhan keseimbangan performa antar kelas pada kedua model relatif setara.

Secara keseluruhan, hasil perbandingan menunjukkan bahwa *IndoBERTweet* masih unggul pada metrik *accuracy*, yang mengindikasikan performa global yang lebih stabil pada keseluruhan data uji. Namun demikian, Skenario 1 menunjukkan keunggulan yang lebih jelas pada metrik kelas toxic, khususnya pada *precision*, *recall*, dan *F1-score*. Hal ini menunjukkan bahwa model *RoBERTa* pada Skenario 1 lebih efektif dalam menangani kelas *toxic*, meskipun performa keseluruhan *accuracy* sedikit lebih rendah.

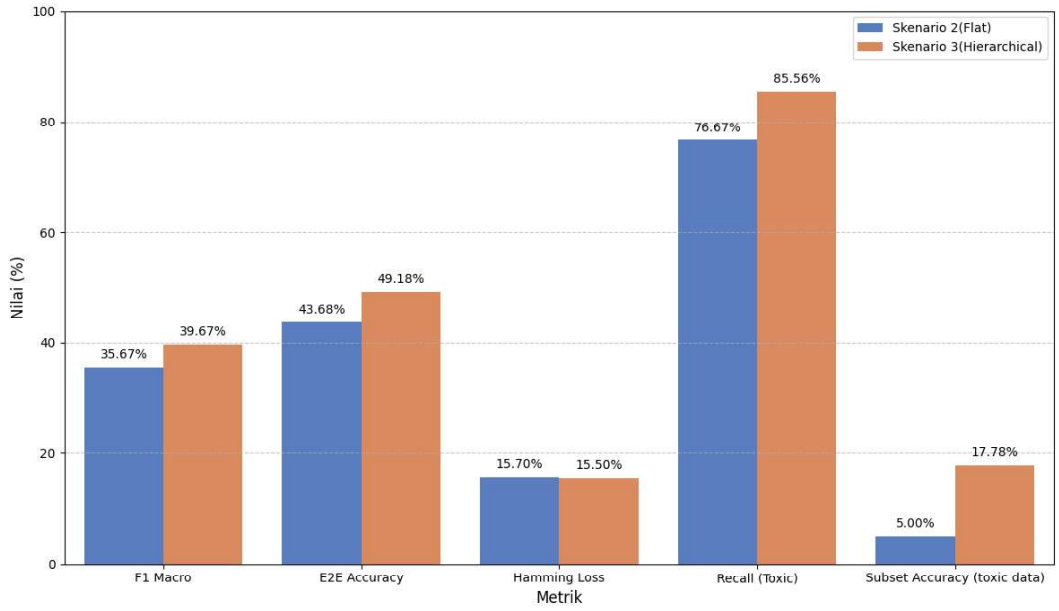
Jika dilihat lebih dalam, nilai *F1-macro* yang hampir setara antara kedua model menunjukkan bahwa Skenario 1 memiliki distribusi performa yang lebih seimbang antar kelas, meskipun tidak didominasi oleh kelas non-toxic. Dalam konteks ini, performa Skenario 1 tidak semata-mata ditentukan oleh keberhasilan klasifikasi kelas mayoritas (non-toxic), tetapi juga oleh kemampuan yang lebih kuat dalam mendeteksi kelas toxic. Dengan demikian, Skenario 1 tidak dapat dikatakan

kalah secara keseluruhan, melainkan menunjukkan karakteristik model yang lebih sensitif terhadap kelas yang lebih penting secara kontekstual.

Berdasarkan hasil tersebut, Skenario 1 dapat dinyatakan sebagai model yang lebih adaptif terhadap deteksi ujaran toksik, meskipun IndoBERTweet masih unggul dalam hal *accuracy* sebagai metrik global. Keunggulan Skenario 1 pada kelas toxic menunjukkan bahwa model ini lebih sesuai untuk skenario deteksi ujaran toksik yang menekankan minimisasi *false negative*. Oleh karena itu, Skenario 1 tetap dapat dikatakan kompetitif dan dalam aspek tertentu bahkan lebih relevan terhadap tujuan utama penelitian dibandingkan model rujukan *IndoBERTweet*.

4.3.2 Perbandingan Skenario *Flat* dan Skenario *Hierarchical*

Perbandingan pada subbab ini dilakukan antara Skenario 2 yang menggunakan pendekatan *flat multi-label classification* dan Skenario 3 yang menggunakan pendekatan *hierarchical classification* dengan evaluasi *end-to-end*. Evaluasi dilakukan menggunakan data uji (*test set*) yang terdiri dari 364 komentar. Pada Skenario 2, model secara langsung memprediksi seluruh label toksisitas dalam satu tahap klasifikasi. Sementara itu, pada Skenario 3 proses klasifikasi dilakukan secara bertingkat, yaitu melalui deteksi *toxic* dan *non-toxic* pada Level 1, kemudian dilanjutkan dengan klasifikasi *multi-label* kategori toksisitas pada Level 2 untuk komentar yang terdeteksi *toxic*.



Gambar 4.8 Perbandingan Skenario *flat* vs Skenario *Hierarchical*

Berdasarkan Gambar 4.2, Skenario 3 menunjukkan performa yang lebih baik dibandingkan Skenario 2 pada hampir seluruh metrik evaluasi. Nilai *F1 Macro* meningkat dari 35,67% menjadi 39,67%, yang menunjukkan bahwa pendekatan *hierarchical classification* mampu menghasilkan keseimbangan *precision* dan *recall* yang lebih baik pada seluruh label toksisitas. Peningkatan ini mengindikasikan bahwa pemisahan proses klasifikasi ke dalam dua tahap membantu model dalam mengenali karakteristik setiap kategori toksisitas secara lebih efektif.

Pada metrik *E2E Accuracy*, Skenario 3 memperoleh nilai 49,18%, lebih tinggi dibandingkan Skenario 2 yang mencapai 43,68%. *E2E Accuracy* merupakan metrik yang ketat karena seluruh label yang diprediksi harus sama persis dengan label sebenarnya. Oleh karena itu, peningkatan sebesar 5,50% menunjukkan bahwa pendekatan bertingkat lebih mampu menghasilkan kombinasi label yang benar

secara keseluruhan. Selain itu, nilai *Hamming Loss* pada Skenario 3 mengalami penurunan dari 15,70% menjadi 15,50%. Nilai *Hamming Loss* yang lebih rendah menunjukkan bahwa jumlah kesalahan prediksi label pada setiap sampel semakin sedikit. Meskipun penurunannya relatif kecil, hasil ini menunjukkan bahwa proses penyaringan pada Level 1 tetap memberikan kontribusi dalam mengurangi kesalahan klasifikasi yang terjadi pada tahap *multi-label*.

Perbandingan kemampuan deteksi komentar *toxic* juga menunjukkan perbedaan yang cukup jelas. Skenario 2 mendapatkan nilai recall *toxic* di 76,67%, sedangkan Skenario 3 berhasil mendapatkan nilai recall *toxic* lebih tinggi di 85,56%. Peningkatan sebesar 8,89% menunjukkan bahwa pendekatan *hierarchical classification* lebih efektif dalam mendeteksi keberadaan toksisitas pada data uji. Namun demikian, perbedaan yang paling signifikan terlihat pada metrik *Subset Accuracy (toxic data)* yang menilai *accuracy* multilabel. Pada Skenario 2 hanya mendapatkan *Subset Accuracy* 5%, dimana hanya 9 dari 180 komentar *toxic* yang seluruh label toksisitasnya berhasil diprediksi secara sempurna. Sebaliknya, Skenario 3 mampu menghasilkan 32 prediksi sempurna (17,78%). Peningkatan sebesar 12,78% menunjukkan bahwa pendekatan bertingkat jauh lebih efektif dalam menentukan kombinasi kategori toksisitas yang tepat pada komentar yang memang bersifat *toxic*.

Secara keseluruhan, hasil pengujian menunjukkan bahwa pendekatan *hierarchical classification* pada Skenario 3 memberikan performa yang lebih baik dibandingkan pendekatan *flat multi-label classification* pada Skenario 2. Meskipun kedua skenario sama-sama mampu melakukan klasifikasi toksisitas, Skenario 3

menunjukkan peningkatan pada seluruh metrik utama dan menghasilkan nilai *Hamming Loss* yang lebih rendah. Oleh karena itu, pendekatan *hierarchical classification* dapat dianggap lebih efektif untuk menangani permasalahan klasifikasi toksisitas bertingkat pada dataset penelitian ini.

3.4 Pembahasan

Pembahasan pada subbab ini menjelaskan interpretasi dari seluruh hasil pengujian yang telah dilakukan pada Skenario 1, Skenario 2, dan Skenario 3. Fokus pembahasan tidak hanya pada nilai metrik yang diperoleh, tetapi juga pada alasan di balik perbedaan performa antar pendekatan yang digunakan. Selain itu, pembahasan ini juga menekankan hubungan antara karakteristik data, arsitektur model, serta strategi klasifikasi yang diterapkan. Dengan demikian, hasil evaluasi tidak hanya dipahami sebagai angka, tetapi juga sebagai representasi dari perilaku model terhadap kompleksitas permasalahan.

Pada Skenario 1, model *RoBERTa* digunakan sebagai classifier biner independen untuk membedakan teks toxic dan non-toxic. Hasil evaluasi menunjukkan bahwa model memiliki performa yang cukup baik dengan *F1-score* sebesar 0,7907 dan *recall* sebesar 0,8558. Nilai *recall* yang tinggi menunjukkan bahwa model cenderung lebih sensitif dalam mendeteksi kelas toxic dibandingkan kelas non-toxic. Hal ini mengindikasikan bahwa model lebih berorientasi pada minimisasi kasus *false negative*, yang penting dalam konteks deteksi ujaran toksik.

Jika dibandingkan dengan Skenario 3 pada Level 1, kedua pendekatan memiliki task yang secara definisi sama, yaitu klasifikasi biner toxic dan non-toxic. Namun demikian, Skenario 3 Level 1 merupakan bagian dari sistem *hierarchical*

classification, sehingga perannya tidak hanya sebagai classifier independen, tetapi juga sebagai tahap penyaringan awal sebelum proses klasifikasi lanjutan. Hasil evaluasi menunjukkan bahwa Skenario 3 Level 1 memiliki performa yang sedikit lebih tinggi dengan *accuracy* sebesar 0,7964 dan *F1-score* sebesar 0,8001.

Perbedaan performa antara Skenario 1 dan Skenario 3 Level 1 tidak hanya disebabkan oleh model, tetapi juga oleh perbedaan konteks optimasi. Skenario 1 dilatih sebagai model independen yang berfokus pada klasifikasi akhir, sedangkan Skenario 3 Level 1 dioptimalkan sebagai bagian dari pipeline sehingga hasil prediksi digunakan untuk menentukan apakah suatu data akan diproses lebih lanjut pada Level 2. Hal ini menyebabkan model pada Skenario 3 lebih terarah dalam membentuk boundary klasifikasi yang stabil untuk mendukung tahap berikutnya dalam sistem.

Pada Skenario 2 yang menggunakan pendekatan *flat multi-label classification*, hasil evaluasi menunjukkan performa yang relatif lebih rendah dibandingkan pendekatan lainnya. Nilai *macro F1* sebesar 0,3207 menunjukkan bahwa model mengalami kesulitan dalam mempelajari seluruh label secara bersamaan. Hal ini disebabkan oleh kompleksitas hubungan antar label serta ketidakseimbangan distribusi data pada masing-masing kategori. Akibatnya, model cenderung lebih baik pada label dengan jumlah data besar seperti toxicity, namun lemah pada label minor seperti *threat_incitement_to_violence*.

Berbeda dengan Skenario 2, Skenario 3 menggunakan pendekatan bertingkat yang membagi proses klasifikasi menjadi Level 1 dan Level 2. Pada Level 2, model hanya memproses data yang telah teridentifikasi sebagai toxic,

sehingga ruang permasalahan menjadi lebih terfokus. Hasil evaluasi menunjukkan bahwa *macro F1* pada Level 2 sebesar 0,5247, yang secara numerik cukup jauh dengan Skenario 2. Namun demikian, pendekatan ini memberikan struktur pembelajaran yang lebih terkontrol karena mengurangi noise dari data non-toxic.

Jika dilihat pada evaluasi *end-to-end*, Skenario 3 menunjukkan keunggulan yang lebih jelas dibandingkan Skenario 2. Nilai *Subset Accuracy* pada Skenario 3 mencapai 49,18%, lebih tinggi dibandingkan Skenario 2 yang hanya sebesar 43,68%. Hal ini menunjukkan bahwa pendekatan *hierarchical* lebih mampu menghasilkan kombinasi label yang sepenuhnya sesuai dengan ground truth. Selain itu, distribusi prediksi pada Skenario 3 juga lebih stabil karena proses klasifikasi dilakukan secara bertahap.

Secara keseluruhan, hasil penelitian menunjukkan bahwa struktur arsitektur model memiliki pengaruh yang signifikan terhadap performa klasifikasi ujaran toksik. Skenario 1 menunjukkan performa yang kuat sebagai classifier biner independen, Skenario 2 menunjukkan keterbatasan dalam menangani kompleksitas multi-label, sedangkan Skenario 3 memberikan keseimbangan terbaik antara deteksi awal dan klasifikasi lanjutan. Dengan demikian, pendekatan *hierarchical classification* terbukti lebih efektif untuk menyelesaikan permasalahan klasifikasi ujaran toksik yang bersifat bertingkat dan tidak seimbang.

Berdasarkan hasil pengujian, skenario 3 menghasilkan performa yang lebih baik dibandingkan skenario lainnya. Pada skenario ini, proses klasifikasi dilakukan secara hierarkis melalui dua tahapan. Tahap pertama bertujuan untuk mengidentifikasi apakah suatu komentar termasuk kategori *toxic* atau *non-toxic*.

Selanjutnya, hanya komentar yang terdeteksi *toxic* yang diproses pada tahap kedua untuk menentukan jenis toksisitasnya. Pendekatan ini memungkinkan model untuk mempelajari karakteristik yang lebih spesifik pada setiap tahap sehingga menghasilkan proses klasifikasi yang lebih terarah.

Konsep tersebut sejalan dengan ajaran Islam yang menekankan pentingnya keteraturan dan penempatan sesuatu sesuai dengan kategorinya. Allah SWT berfirman dalam QS. Al-Qamar ayat 49:

إِنَّا كُلَّ شَيْءٍ خَلَقْنَٰهُ بِقَدَرٍ ﴿٤٩﴾

"Sesungguhnya Kami menciptakan segala sesuatu menurut ukuran." (QS. Al-Qamar: 49)

Ayat tersebut menjelaskan bahwa segala sesuatu yang Allah SWT ciptakan memiliki ukuran, ketetapan, dan keteraturan tertentu. Berdasarkan penjelasan dalam Tafsir Kementerian Agama Republik Indonesia, ayat ini menerangkan bahwa seluruh kejadian di alam semesta berlangsung berdasarkan ketentuan dan aturan yang telah ditetapkan oleh Allah SWT, sehingga tidak terdapat sesuatu yang terjadi secara sia-sia atau tanpa perencanaan (Kementerian Agama Republik Indonesia, 2019). Prinsip keteraturan tersebut menunjukkan bahwa setiap sesuatu memiliki karakteristik dan ketentuan yang membedakannya.

Konsep tersebut memiliki keterkaitan dengan pendekatan *hierarchical classification* yang diterapkan pada skenario 3. Model tidak langsung menentukan seluruh jenis toksisitas dalam satu tahap, tetapi terlebih dahulu mengidentifikasi kategori utama, yaitu komentar *toxic* dan *non-toxic*. Setelah komentar teridentifikasi sebagai *toxic*, model melakukan klasifikasi lanjutan untuk

menentukan jenis toksisitas berdasarkan karakteristik yang terkandung dalam komentar tersebut.

Pendekatan ini menunjukkan bahwa proses pengelompokan berdasarkan indikator tertentu dapat membantu menghasilkan klasifikasi yang lebih terarah. Sebagaimana hadis menjelaskan suatu karakteristik melalui tanda-tanda yang dapat diamati, model pada penelitian ini juga memanfaatkan pola-pola yang terdapat pada data untuk mengenali kategori yang sesuai. Oleh karena itu, struktur klasifikasi yang dilakukan secara bertahap memungkinkan model memahami hubungan antara kategori umum dan kategori yang lebih spesifik sehingga menghasilkan performa klasifikasi yang lebih baik.

Selain konsep keteraturan dalam pengelompokan, pendekatan klasifikasi bertahap juga memiliki kesamaan dengan cara Islam menjelaskan karakteristik suatu golongan berdasarkan ciri-ciri tertentu. Rasulullah SAW bersabda:

آيَةُ الْمُنَافِقِ ثَلَاثٌ: إِذَا حَدَّثَ كَذَبَ، وَإِذَا وَعَدَ أَخْلَفَ، وَإِذَا اؤْتُمِنَ خَانَ

"Tanda-tanda orang munafik ada tiga: apabila berbicara ia berdusta, apabila berjanji ia mengingkari, dan apabila diberi amanah ia berkhianat." (HR. Bukhari dan Muslim)

Dalam penjelasannya terhadap hadits tersebut, Imam an-Nawawi dalam *Syarh Shahih Muslim* menerangkan bahwa sifat-sifat yang disebutkan oleh Rasulullah SAW merupakan indikator yang dapat digunakan untuk mengenali karakteristik kemunafikan dalam perilaku (*nifaq amali*). Dengan kata lain, suatu karakteristik dapat dikenali melalui tanda-tanda atau ciri-ciri tertentu yang menjadi pembeda dengan karakteristik lainnya (An-Nawawi, 2010).

Dalam konteks penelitian ini, prinsip keteraturan tersebut tercermin pada pendekatan hierarchical classification yang digunakan pada skenario 3. Model tidak langsung mengklasifikasikan seluruh jenis toksisitas secara bersamaan, melainkan melakukan pengelompokan secara bertahap sesuai karakteristik data. Komentar terlebih dahulu dipisahkan berdasarkan status *toxic* dan *non-toxic*, kemudian komentar yang termasuk *toxic* diklasifikasikan lebih lanjut ke dalam jenis toksisitas yang sesuai. Dengan struktur seperti ini, model dapat lebih fokus mempelajari pola yang relevan pada setiap tahap klasifikasi.

Dari sisi machine learning, pemisahan proses klasifikasi ke dalam beberapa tingkatan juga membantu mengurangi kompleksitas tugas yang harus dipelajari model dalam satu waktu. Model level kedua hanya memproses komentar yang telah teridentifikasi *toxic*, sehingga proses pembelajaran menjadi lebih terarah dibandingkan ketika seluruh komentar langsung diklasifikasikan ke berbagai jenis toksisitas secara bersamaan. Hal ini sejalan dengan hasil penelitian yang menunjukkan bahwa skenario 3 memperoleh nilai evaluasi yang lebih tinggi dibandingkan skenario lainnya.

Oleh karena itu, keunggulan skenario 3 tidak hanya menunjukkan efektivitas pendekatan hierarkis dari sisi teknis, tetapi juga mencerminkan nilai keteraturan dan pengelompokan yang diajarkan dalam Islam sebagaimana terkandung dalam QS. Al-Qamar ayat 49, bahwa segala sesuatu diciptakan dengan ukuran dan susunan yang teratur.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian, model *FacebookAI/xlm-roberta-base* mampu digunakan untuk melakukan klasifikasi ujaran toksik berbahasa Indonesia pada tiga skenario pengujian, yaitu klasifikasi biner, *flat multi-label classification*, dan *hierarchical multi-label classification*. Pada skenario klasifikasi biner, model menunjukkan kemampuan yang baik dalam membedakan komentar *toxic* dan *non-toxic* dengan nilai *precision*, *recall*, *F1-score*, dan *accuracy* yang tinggi. Pada skenario *flat multi-label classification*, model mampu mengidentifikasi lebih dari satu kategori toksisitas dalam satu komentar secara simultan, meskipun masih menghadapi tantangan akibat ketidakseimbangan distribusi label dan kompleksitas hubungan antar kategori toksisitas. Hasil evaluasi menggunakan metrik *precision*, *recall*, *F1-score*, *Macro-F1*, *accuracy*, *Subset Accuracy*, dan *Hamming Loss* menunjukkan bahwa model dapat digunakan untuk menangani tugas klasifikasi ujaran toksik bertingkat pada media sosial.

Berdasarkan perbandingan performa ketiga skenario, pendekatan *hierarchical multi-label classification* memberikan hasil yang lebih baik dibandingkan *flat multi-label classification*. Pada evaluasi *end-to-end*, pendekatan *hierarchical classification* memperoleh nilai *Macro-F1* sebesar 39,67%, *E2E Accuracy* sebesar 49,18%, dan *Hamming Loss* sebesar 15,5%, sedangkan pendekatan *flat multi-label classification* memperoleh nilai *Macro-F1* sebesar 35,67%, *accuracy* sebesar 43,68%, dan *Hamming Loss* sebesar 15,7%. Selain itu,

hierarchical classification menghasilkan *Subset Accuracy* pada data *toxic* sebesar 17,78%, lebih tinggi dibandingkan *flat multi-label classification* yang hanya mencapai 5%. Hasil tersebut menunjukkan bahwa pemisahan proses klasifikasi ke dalam dua tahap, yaitu deteksi *toxic* dan *non-toxic* pada *Level 1* serta klasifikasi kategori toksisitas pada *Level 2*, mampu meningkatkan kualitas prediksi kombinasi label dan mengurangi kesalahan klasifikasi, sehingga berdasarkan hasil evaluasi pada *hold-out test set*, pendekatan *hierarchical classification* menunjukkan performa lebih baik dibandingkan *flat multi-label classification* pada *F1 Macro, E2E Accuracy,* dan *Subset Accuracy* data toxic. Namun, peningkatan pada *Hamming Loss* relatif kecil, sehingga keunggulan hierarchical lebih terlihat pada ketepatan kombinasi label dibandingkan pada penurunan kesalahan per-label.

5.2 Saran

Berdasarkan hasil penelitian yang telah dilakukan, penelitian selanjutnya disarankan untuk menggunakan jumlah data yang lebih besar serta distribusi label yang lebih seimbang agar model dapat mempelajari karakteristik setiap kategori toksisitas secara lebih optimal, khususnya pada label dengan jumlah data terbatas seperti *threat* dan *identity attack*. Selain itu, penelitian berikutnya dapat mengeksplorasi berbagai metode penanganan ketidakseimbangan data, seperti *oversampling*, *data augmentation*, maupun *class weighting*, guna meningkatkan kemampuan model dalam mengklasifikasikan label minoritas dan meningkatkan performa klasifikasi secara keseluruhan.

Selain pengembangan pada aspek data, penelitian selanjutnya juga dapat melakukan perbandingan performa model *FacebookAI/xlm-roberta-base* dengan

model *Transformer* lainnya, seperti *IndoBERT*, *IndoBERTweet*, dan *mBERT*, untuk mengetahui model yang paling sesuai dalam tugas klasifikasi ujaran toksik *multi-label* berbahasa Indonesia. Penelitian berikutnya juga dapat mengembangkan struktur *hierarchical classification* yang lebih kompleks, misalnya dengan menambahkan level hierarki yang lebih rinci atau menerapkan pendekatan yang mempertimbangkan hubungan antarlabel, sehingga karakteristik ujaran toksik dapat direpresentasikan secara lebih baik dan menghasilkan kualitas klasifikasi yang lebih tinggi.

DAFTAR PUSTAKA

Allgaier, J., & Pryss, R. (2024). Practical approaches in evaluating validation and biases of machine learning applied to mobile health studies. *Communications Medicine*, *4*(1), 76. https://doi.org/10.1038/s43856-024-00468-0

Amin, K., Dziqie, M., Alfarauqi, A., & Khatimah, K. (2018). *SOCIAL MEDIA, CYBER HATE, AND RACISM*. https://doi.org/10.23917/komuniti.v10i1.5613

Azhari, A. A., Sibaroni, Y., & Prasetiyowati, S. S. (2023). Detection of Indonesian Hate Speech in the Comments Column of Indone-sian Artists' Instagram Using the RoBERTa Method. *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, *8*(3), 764–773. https://doi.org/10.29100/jipi.v8i3.3898

Bate, A. P. (2024). Melawan Toksisitas di Media Sosial melalui Edukasi Etika Siber (Pengabdian Masyarakat pada SMKN 1 Bayah, Lebak). *Inovasi Jurnal Pengabdian Masyarakat*, *2*(1), 133–140. https://doi.org/10.54082/ijpm.395

Cohen, J., & Fung, A. (2023). Democratic responsibility in the digital public sphere. *Constellations*, *30*(1), 92–97. https://doi.org/https://doi.org/10.1111/1467-8675.12670

Das, M., & Balke, W.-T. (2024). Toximatics: Towards Understanding Toxicity in Real-Life Social Situations. Dalam T. Kawahara, V. Demberg, S. Ultes, K. Inoue, S. Mehri, D. Howcroft, & K. Komatani (Ed.), *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (hlm. 770–785). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.sigdial-1.65

Findawati, Y., Raharjo, A. B., Navastara, A., Yonathan, V., Yatestha, A., & Purwitasari, D. (2023). *Multi-label Aspect Dangerous Speech Classification Using Keyword-Driven Ensemble Classifier on Imbalanced Data*. www.joiv.org/index.php/joiv

Fortuna, P., & Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys (CSUR)*, *51*(4).

Handayani, T. P., & Hilmansyah Gani. (2023). Enhancing Multi-Label Hate Speech and Abusive Language Detection on Indonesian Twitter Using Recurrent Neural Networks with Hyperparameter Tuning. *Jurnal Ilmiah Teknik Mesin, Elektro dan Komputer*, *3*(3), 601–612. https://doi.org/10.51903/juritek.v3i3.3022

Ibrohim, M. O., & Budi, I. (2019). *Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter*. https://www.komnasham.go.id/index.php/

Koto, F., Lau, J. H., & Baldwin, T. (2021). IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization. Dalam M.-F. Moens, X. Huang, L. Specia, & S. W. Yih (Ed.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (hlm. 10660–10668). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.833

Koto Jey Han Lau Timothy Baldwin, F. (2021). *INDOBERTWEET: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization*. https://huggingface.co/huseinzol05/

Kusuma, A. P., & Rahmawati, N. K. (2025). Toxic language detection on social media: a critical linguistic approach to online hate speech. Dalam *Journal of Life-Span Psychology, Linguistics, and Media Studies* (Vol. 01, Nomor 01). https://www.jllm.spdfharmony.com/index.php/jllm

LOPEZ, A. B., Pastor-Galindo, J., & Ruipérez-Valiente, J. A. (2024). *Exploring the topics, sentiments and hate speech in the Spanish information environment*. https://arxiv.org/abs/2409.12658

Novry, H., Paninggiran, K., Rusadi, U., Wicaksana, D. A., & Aulia, W. M. (2025). *BIJMT : Brilliant International Journal Of Management And Tourism Media Spatialization in Indonesia's Digital Political Econ-omy: A Case Study of Platform X through Vincent Mosco's Framework*. https://doi.org/10.55606/bijmt.v5i2.45072

Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., & Androutsopoulos, I. (2020). *Toxicity Detection: Does Context Really Matter?* http://arxiv.org/abs/2006.00998

Sahu, P., & S R, Dr. N. (2024). Enhancing Sentiment Analysis using BERT-Hybrid Model for Detection of Irony and Sarcasm in Code-Mixed Social Media. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, *08*(01), 1–13. https://doi.org/10.55041/IJSREM28163

Santoso, D. H., Aziz, J., Pawito, Utari, P., & Kartono, D. T. (2020). Populism in new media: The online presidential campaign discourse in Indonesia. *GEMA Online Journal of Language Studies*, *20*(2), 115–133. https://doi.org/10.17576/gema-2020-2002-07

Sukidin, Hudha, C., & Basrowi. (2025). Shaping democracy in Indonesia: The influence of multicultural attitudes and social media activity on participation in public discourse and attitudes toward democracy. *Social Sciences and Humanities Open*, *11*. https://doi.org/10.1016/j.ssaho.2025.101440

Susanto, L., Wijanarko, M. I., Pratama, P. A., Hong, T., Idris, I., Aji, A. F., & Wijaya, D. (2025). *IndoToxic2024: A Demographically-Enriched Dataset of Hate Speech and Toxicity Types for Indonesian Language*. http://arxiv.org/abs/2406.19349

Susanto, L., Wijanarko, M., Pratama, P., Tang, Z., Akyas, F., Hong, T., Idris, I., Aji, A., & Wijaya, D. (2025). *A Multi-Labeled Dataset for Indonesian Discourse: Examining Toxicity, Polarization, and Demographics Information*. http://arxiv.org/abs/2503.00417

Van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). *Challenges for Toxic Comment Classification: An In-Depth Error Analysis*. http://www.perspectiveapi.com/

Wang, Z., Cheng, J., Cui, C., & Yu, C. (2023). *Implementing BERT and fine-tuned RobertA to detect AI generated news by ChatGPT*.

Wong, T.-T., & Yeh, P.-Y. (2020). Reliable Accuracy Estimates from k-Fold Cross Validation. *IEEE Transactions on Knowledge and Data Engineering*, *32*(8), 1586–1594. https://doi.org/10.1109/TKDE.2019.2912815

Xu, L., Teng, S., Zhao, R., Guo, J., Xiao, C., Jiang, D., & Ren, B. (2021). *Hierarchical Multi-label Text Classification with Horizontal and Vertical Category Correlations*.

Xu, Y., & Goodacre, R. (2018). On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Journal of Analysis and Testing*, *2*(3), 249–262. https://doi.org/10.1007/s41664-018-0068-2

Zhou, G., Wang, H., Jin, D., Wang, W., Jiang, S., Tang, R., & Chen, X. (2025). A Toxic Euphemism Detection framework for online social network based on Semantic Contrastive Learning and dual channel knowledge augmentation. *Information Processing & Management*, *62*(4), 104143. https://doi.org/10.1016/J.IPM.2025.104143