

**KLASIFIKASI HIERARKIS UJARAN KEBENCIAN MULTI-LABEL DENGAN
INDOBERTWEET DAN REGRESI ORDINAL**

SKRIPSI

Oleh:
SITI WASI'ATUL MAGHFIROH
NIM. 220605110035



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

**KLASIFIKASI HIERARKIS UJARAN KEBENCIAN MULTI-LABEL
DENGAN INDOBERTWEET DAN REGRESI ORDINAL**

SKRIPSI

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

Oleh:
SITI WASI'ATUL MAGHFIROH
NIM. 220605110035

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

HALAMAN PERSETUJUAN

KLASIFIKASI HIERARKIS UJARAN KEBENCIAN MULTI-LABEL DENGAN INDOBERTWEET DAN REGRESI ORDINAL

SKRIPSI

Oleh:
SITI WASI'ATUL MAGHFIROH
NIM. 220605110035

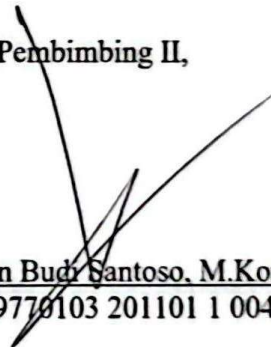
Telah Diperiksa dan Disetujui untuk Diuji:
Tanggal: 30 April 2026

Pembimbing I,



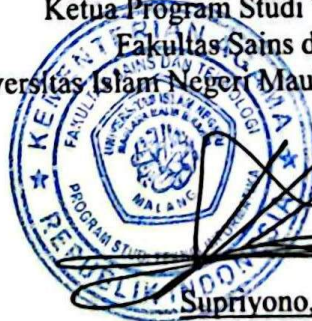
Khadijah Fahmi Hayati Holle, M.Kom
NIP. 19900626 202203 2 002

Pembimbing II,



Dr. Irwan Budi Santoso, M.Kom
NIP. 19770103 201101 1 004

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 KEBENCIAN MULTI-LABEL DENGAN INDOBERTWEET DAN REGRESI ORDINAL

SKRIPSI

SITI WASI'ATUL MAGHFIROH
NIM. 220605110035

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 30 April 2026

Susunan Dewan Penguji

Ketua Penguji : Dr. M. Amin Hariyadi, M.T
NIP. 19670118 200501 1 001

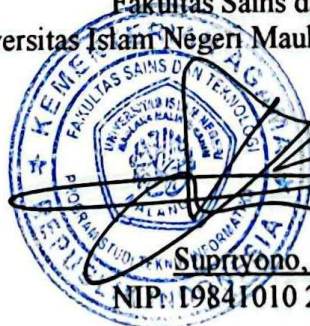
Anggota Penguji I : Okta Qomaruddin Aziz, M.Kom
NIP. 19911019 201903 1 013

Anggota Penguji II : Khadijah Fahmi Hayati Holle, M.Kom
NIP. 19900626 202203 2 002

Anggota Penguji III : Dr. Irwan Budi Santoso, M.Kom
NIP. 19770103 201101 1 004

()
()
()
()

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 : Siti Wasi'atul Maghfiroh
NIM : 220605110035
Fakultas / Program Studi : Sains dan Teknologi / Teknik Informatika
Judul Skripsi : Klasifikasi Hierarkis Ujaran Kebencian
Multi-label Dengan IndoBERTweet Dan Regresi
Ordinal

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, 08 Juni 2026
Yang membuat pernyataan,



Siti Wasi'atul Maghfiroh
NIM.220605110035

MOTTO

... *“If you want something work for it!”* ...

HALAMAN PERSEMBAHAN

Segala puji bagi Allah SWT yang telah memberikan rahmat dan hidayah-Nya,
sehingga penulis dapat menyelesaikan tugas akhir ini.

Dengan penuh rasa syukur skripsi ini saya persembahkan kepada:

Orang tua saya,

Yang selalu mendoakan, menguatkan tanpa banyak kata, dan menjadi alasan utama saya untuk terus melangkah hingga menyelesaikan apa yang saya mulai.

Kakak saya,

yang selalu memberikan dukungan dan pengorbanan yang tidak selalu terlihat, namun sangat berarti dalam perjalanan saya.

Adik-Adik saya,

yang menjadi penyemangat saya untuk tidak menyerah ketika dan memberikan warna dalam proses penyelesaian tugas akhir ini.

Para Dosen,

yang telah membimbing dengan sabar, memberikan ilmu, arahan, dan dukungan sehingga penulis dapat menyelesaikan tugas akhir ini dengan baik.

Teman-Teman Angkatan 2022 Teknik Informatika (*infinity*)

Yang telah menjadi bagian dari perjalanan ini, berbagi cerita, perjuangan, dan tawa sehingga masa perkuliahan lebih terasa ringan dan penuh warna.

KATA PENGANTAR

Assalamu 'alaikum Wr. Wb.

Alhamdulillah, segala puji dan rasa syukur penulis panjatkan ke hadirat Allah Subhanahu Wa Ta'ala atas rahmat, karunia, dan hidayah-Nya, sehingga penulis dapat menyelesaikan skripsi yang berjudul “Klasifikasi Hierarkis Ujaran Kebencian Multi-Label Dengan IndoBERTweet Dan Regresi Ordinal”. Sholawat serta salam tetap tercurahkan kepada junjungan kita Nabi Muhammad Shallallahu Alaihi Wasallam, yang menjadi suri tauladan dan semoga kita mendapat syafaat di akhir kelak, Aamiin.

Dalam penyusunan skripsi ini, penulis menyadari bahwa banyak pihak telah memberikan doa dan dukungan. Oleh karena itu, penulis mengucapkan terima kasih yang sebesar-besarnya kepada:

1. Prof. Dr. Hj. Ilfi Nur Diana, M.Si., CAHRM., CRMP., selaku Rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Dr. Agus Mulyono, M.Kes., selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Supriyono, M.Kom., selaku Ketua Program Studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Khadijah Fahmi Hayati Holle, M.Kom., selaku Dosen Pembimbing 1, yang sudah membimbing saya dari awal, dengan sabar, selalu memberi saya motivasi ketika saya mulai lelah, dan memberi afirmasi positif ketika saya ketakutan. Terima kasih banyak atas waktu untuk berdiskusi bersama ketika

saya kebingungan, beliau juga senantiasa memberi semangat dan do'a sehingga saya bisa menyelesaikan penelitian ini. Terima kasih banyak ibu semoga apa yang ibu berikan bisa berkah kepada semuanya dan ibu selalu dimudahkan dalam segala hal oleh Allah.

5. Dr. Irwan Budi Santoso, M.Kom., selaku Dosen Pembimbing 2, terima kasih banyak bapak telah membimbing saya dan membantu penulisan saya agar penelitian ini lebih terarah dan bisa dipertanggung jawabkan.
6. Dr. M. Amin Hariyadi, M.T., selaku ketua penguji, yang telah memberikan ilmu dan saran serta nasihat terhadap penulis sehingga bisa menyusun skripsi ini dengan baik dan lebih terstruktur.
7. Okta Qomaruddin Aziz, M.Kom., Selaku anggota penguji 1 yang telah memberikan arahan, bimbingan, serta solusi selama proses penyusunan tugas akhir ketika penulis mengalami kebingungan. Beliau juga banyak membagikan pengalaman yang bermanfaat bagi kami semua sehingga dapat menjadi pembelajaran bagi kami ke depannya. Penulis mengucapkan terima kasih atas segala bantuan yang telah diberikan, semoga kebaikan beliau dibalas dengan kemudahan dalam setiap urusan serta keberkahan dalam setiap langkah.
8. Nia Faricha, S.Si., selaku Admin Program Studi Teknik Informatika, yang dengan penuh kesabaran dan ketulusannya membantu penulis dalam berbagai urusan administrasi dari awal perkuliahan sampai akhirnya bisa menyelesaikan skripsi ini. Terima kasih banyak mba nia atas informasi dan arahannya selama ini, semoga selalu dimudahkan dalam segala urusannya.

9. Seluruh dosen dan staf Program Studi Teknik Informatika yang telah memberikan ilmu, pengalaman, bimbingan, serta dukungan yang berharga, dan membantu kelancaran proses akademik maupun administrasi selama penulis menempuh pendidikan hingga menyelesaikan studi ini.
10. Cinta Pertama Anak Perempuannya, Bapak Penulis. Terima kasih atas segala cinta, perjuangan, pengorbanan, dan doa yang telah diberikan selama ini, beliau selalu mengusahakan agar keluarganya tidak pernah merasa kekurangan, mendukung setiap mimpi anak-anaknya, serta mengajarkan tentang rasa “*Syukur*” dalam setiap keadaan. Beliau juga menanamkan nilai-nilai akhlak dan agama yang menjadi bekal berharga dalam kehidupan penulis. Berkat kerja keras, kasih sayang, dan didikan beliau, penulis dapat menempuh pendidikan hingga berada pada tahap penyelesaian tugas akhir ini. Beliau juga selalu memberikan keyakinan untuk berani melangkah dalam setiap keadaan seperti halnya nasihat beliau yang selalu penulis ingat “Keyakinan tidak dapat dihilangkan dengan keraguan”. Terima kasih untuk segalanya, untuk cinta, perjuangan, dan doa yang tidak pernah putus. Baik-baik di sana, bapak. Peluk jauh dan rindu yang tak pernah usai dari penulis untuk bapak.
11. Untuk Wanita Tangguh, Ibu Penulis. Terima kasih banyak ibu, banyak hal yang sudah beliau korbankan untuk penulis, beliau selalu mendoakan, mengusahakan, menemani penulis dari awal sampai sekarang. Beliau juga senantiasa mendoakan, mengusahakan yang terbaik, serta menemani setiap langkah penulis dari awal perkuliahan hingga tahap penyelesaian tugas

akhir ini. Berbagai nasihat yang diberikan, terutama untuk mengutamakan pendidikan dan fokus pada cita-cita yang ingin dicapai, menjadi penguat bagi penulis dalam menjalani setiap proses kehidupan. Dengan penuh kasih sayang, Ibu selalu berupaya memberikan kebahagiaan dan yang terbaik bagi anak-anaknya. Semoga Ibu senantiasa diberikan kesehatan, kebahagiaan, keberkahan, serta dikelilingi oleh orang-orang baik, sebagaimana doa-doa yang tidak pernah putus beliau panjatkan untuk anak-anaknya.

12. Untuk Kakak Penulis, terima kasih telah menjadi kakak yang kuat dan selalu mengupayakan segala hal agar penulis tidak merasa kekurangan. Terima kasih karena selalu hadir untuk penulis, membantu di saat kesulitan, serta memberikan dukungan terhadap setiap mimpi dan langkah yang ingin penulis capai. Kebaikan, perhatian, dan pengorbanan yang telah diberikan menjadi salah satu alasan penulis mampu bertahan hingga berada di titik ini. Semoga kakak selalu dikelilingi oleh hal-hal baik, diberikan kesehatan, kebahagiaan, dan kemudahan dalam setiap urusan. Penulis juga berharap segala mimpi, harapan, dan cita-cita yang kakak perjuangkan dapat terwujud sebagaimana kakak selalu mendukung mimpi-mimpi penuli

13. Untuk Adik-Adik Penulis, terima kasih sudah jadi penyemangat dan memberi warna dalam hidup penulis, sekaligus jadi alasan untuk segera menyelesaikan tugas akhir ini. Semoga kalian selalu bahagia dan terus tumbuh menjadi pribadi yang baik.

14. Untuk Pak Iwan Sekeluarga, terima kasih atas nasihat dan sarannya sehingga penulis bisa menyelesaikan tugas akhir ini dengan baik, terima

kasih banyak sudah menjadi teman baik bapak, dan selalu memberi semangat dengan menceritakan pengalaman beliau yang sangat bermanfaat.

15. Untuk Ria Nurul F, saudara perempuan penulis, terima kasih banyak sudah hadir dan membantu penulis ketika kesusahan, dan menjadi kakak perempuan yang kuat, baik, dan berani untuk penulis.
16. Untuk Wasik Mubarakah, sahabat penulis sejak MTs. Terima kasih atas dukungan, semangat, dan kebersamaan yang telah diberikan selama ini. Penulis banyak belajar dari pengalaman serta ketangguhan yang dimiliki. Penulis bersyukur dapat bertemu dan bertumbuh bersama Wasik hingga saat ini.
17. Untuk Ustadzah Romlah, terima kasih banyak telah mengajarkan penulis mengaji dari tidak bisa menjadi bisa, serta selalu memberikan semangat dan doa. Terima kasih ustadzah, semoga ilmu yang diberikan menjadi amal jariyah dan bermanfaat untuk banyak orang.
18. Untuk Mba Mila, Mba Azza, Mba Rizqoh, yang sudah penulis anggap menjadi saudara perempuannya. Terima kasih banyak sudah membantu, mendengarkan keluh kesah, dan berbagi hal menarik bersama, semoga kalian selalu dipermudah dan selalu dikelilingi hal-hal baik.
19. Untuk Mirza, Mba Neha, dan Utia, terimakasih sudah menemani penulis dari satu asrama sampai satu kampus. selalu memberi kehangatan dan kasih sayangnya kepada penulis layaknya saudara sendiri. Bahagia selalu yaa.
20. Untuk Lailatul Ilmi Silviana, yang sudah menemani penulis dan berjuang bersama sampai akhirnya kita bisa menyelesaikan ini bersama. Terima kasih

atas semua kebaikan dan bantuannya, semoga selalu dikelilingi hal-hal yang baik dan bahagia selalu sill.

21. Untuk Ais terima kasih sudah menemani dan membantu dari semester awal sampai akhir ini, dan selalu memberi semangat sehingga penulis dapat menyelesaikan skripsi ini.
22. Untuk Anis Lailatul Muhimmah, terima kasih sudah jadi *pathner in crime* selama ini. walaupun sering berantem tapi ada kalanya kita saling membantu dan memberi semangat, selalu bertukar cerita tentang mimpi kita dan saling menguatkan satu sama lain terima kasih banyak ya anis, bahagia selalu.
23. Teman teman seperjuang di PPTQ Ulin Nuha (Putri, Nafis, Azil, Rara, Kayla, Sofia, Luki, Salma, Risa), terima kasih banyak sudah menemani dan berjuang bersama penulis sehingga penulis bisa menyelesaikan skripsi ini tanpa rasa kesepian.
24. Untuk Ibu Khusnul Khatimah, yang sudah mengajari saya memasak dan merawat saya selama disini. Terima kasih karena sudah begitu baik dan perhatian kepada penulis, sehingga penulis tidak merasa sendiri selama jauh dari orang tua. Kehangatan ibu menjadi pengobat saat rasa rindu kepada orang tua datang. Semoga ibu selalu diberi kesehatan dan kebahagiaan.
25. *Feniks Generation*, terima kasih banyak sudah menjadi bagian dari keluarga alumni tersebut, kehangatan dan saling support dalam setiap hal merupakan hal yang sangat saya syukuri selama ini.
26. Teman seperjuangan satu dosen bimbingan Ela, dan Ridiy terima kasih banyak sudah membantu, menemani dan memberikan *support* satu sama

lain ketika bimbingan, sehingga penulis bisa lebih berani dalam menyelesaikan skripsi ini.

27. Teman-Teman “*Infinity*” Teknik Informatika 2022, yang telah menjadi bagian dari perjalanan penulis dari awal sampai akhirnya dapat menyelesaikan skripsi ini. Terima kasih untuk kebersamaan, cerita, perjuangan, dan tawa yang membuat masa perkuliahan terasa lebih ringan dan berwarna.

28. Untuk Semua Pihak yang tidak bisa saya sebutkan satu persatu, penulis ucapkan banyak terima kasih sudah membantu, memberikan semangat sehingga penulis bisa menyelesaikan skripsi ini dengan baik, Bahagia selalu semuanya.

29. Terakhir, untuk diri penulis sendiri. Terima kasih karena telah berhasil melewati berbagai fase kehidupan, termasuk masa-masa yang penuh ketidakpastian dan tantangan. Terima kasih karena tetap bertahan, tetap kuat ketika badai datang, dan tetap memiliki mimpi besar serta harapan untuk masa depan. Terima kasih untuk setiap langkah, perjuangan, air mata, dan usaha yang telah mengantarkan penulis sampai pada titik ini. Tetaplah menjadi perempuan yang apa adanya, yang mampu menebarkan kebaikan dan kebahagiaan bagi orang-orang di sekitarnya. Tetaplah menjadi perempuan yang berani menghadapi masa depan, kuat dalam menghadapi setiap ujian, namun tetap memiliki hati yang lembut dan penuh empati. Penulis tidak memiliki banyak kata untuk menggambarkan panjangnya perjalanan ini. Semoga diri ini dapat terus bertumbuh, memberikan manfaat

bagi banyak orang, dan menjadi pribadi yang lebih baik setiap harinya.

Terima kasih untuk segala hal yang telah dilewati. Bahagia selalu, ya.

Penulis memahami bahwa penyusunan skripsi ini masih belum sepenuhnya sempurna. Oleh sebab itu, penulis berharap adanya masukan yang dapat menjadi bahan pembelajaran dan pengembangan di masa yang akan datang. Semoga hasil penelitian ini dapat memberikan nilai positif, wawasan, dan manfaat bagi siapa pun yang membacanya. Sekian dan terima kasih.

Wassalamu'alaikum Wr, Wb.

Malang, 08 Juni 2026

Penulis

DAFTAR ISI

HALAMAN PERSETUJUAN	iii
HALAMAN PENGESAHAN.....	iv
PERNYATAAN KEASLIAN TULISAN	v
MOTTO	vi
HALAMAN PERSEMBAHAN	vii
KATA PENGANTAR.....	viii
DAFTAR ISI.....	xvi
DAFTAR GAMBAR.....	xviii
DAFTAR TABEL	xix
ABSTRAK	xx
ABSTRACT	xxi
مستخلص البحث.....	xxii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	5
1.3 Batasan Masalah	5
1.4 Tujuan Penelitian	5
1.5 Manfaat Penelitian	6
BAB II STUDI PUSTAKA	7
2.1 Penelitian Terkait	7
2.2 Ujaran kebencian di Media Sosial	10
2.3 IndoBERTweet	11
2.4 Klasifikasi Biner.....	14
2.5 Klasifikasi Multi-Label	15
2.6 Tingkat Keparahan Menggunakan Regresi Ordinal.....	17
2.6.1 Konsep Dasar Regresi Ordinal.....	18
2.6.2 <i>Conditional Ordinal Regression for Neural Networks</i>	20
BAB III METODE PENELITIAN	23
3.1 Desain Penelitian.....	23
3.2 Desain Sistem.....	24
3.3 Dataset.....	26
3.4 <i>Preprocessing</i>	34
3.4.1 <i>Text Cleaning</i>	35
3.4.2 <i>Handling Class Imbalance</i>	35
3.5 <i>Split Data</i>	37
3.6 <i>Feature Extraction</i>	37
3.7 Pemodelan	46
3.7.1 Tahap 1: Klasifikasi Biner	47
3.7.2 Tahap 2: Kategori Multi-label	48
3.7.3 Tahap 3: Tingkat Keparahan (<i>Severity level</i>).....	50
3.8 Evaluasi Model.....	53
3.8.1 Evaluasi Klasifikasi	53
3.8.2 Evaluasi Klasifikasi Multi-label	54

3.8.3 Evaluasi Tingkat Keparahan	55
3.8.4 Evaluasi E2E (<i>End to End</i>)	56
3.9 Skenario Uji Coba	58
BAB IV HASIL DAN PEMBAHASAN	59
4.1 Hasil Uji Coba	59
4.1.1 Hasil Uji Skenario 1	59
4.1.2 Hasil Uji Skenario 2	83
4.1.3 Hasil Uji Skenario 3	107
4.2 Pembahasan	111
4.2.1 Analisis Performa Model Terbaik	111
4.2.2 Analisis Pengaruh <i>Threshold</i>	125
4.2.3 Analisis Pengaruh <i>Balancing</i> Data Pada Model	126
4.2.4 Analisis Kesalahan (<i>Error Analysis</i>)	128
BAB V KESIMPULAN DAN SARAN	137
5.1 Kesimpulan	137
5.2 Saran	138
DAFTAR PUSTAKA	

DAFTAR GAMBAR

Gambar 3.1 Desain Penelitian.....	23
Gambar 3.2 Desain Sistem.....	25
Gambar 3.3 Isi Dataset.....	28
Gambar 3.4 Visualisasi EDA HS dan Non-HS.....	29
Gambar 3.5 Visualisasi EDA <i>Abusive</i> dan <i>Non Abusive</i>	30
Gambar 3.6 Visualisasi EDA Multi-label Hate Speech.....	30
Gambar 3.7 Jumlah Label Dalam 1 Tweet.....	31
Gambar 3.8 Visualisasi EDA Tingkat Keparahan (<i>Severity Level</i>)	34
Gambar 3.9 Arsitektur <i>Fine-Tuning</i> IndoBERTweet.....	39
Gambar 3.10 <i>Flowchart</i> Proses Pelatihan Model IndoBERTweet	46
Gambar 3.11 Alur klasifikasi biner.....	47
Gambar 3.12 Alur Klasifikasi Multi-Label.....	49
Gambar 3.13 Kategori Tingkat Keparahan	50
Gambar 3.14 Arsitektur Model dan Mekanisme Klasifikasi Menggunakan CORN	51
Gambar 4.1 <i>Confusion Matrix</i> Klasifikasi Biner HS dan NON-HS	65
Gambar 4.2 <i>Confusion Matrix</i> Tingkat Keparahan Pada Tahap 3 Skenario 1.....	77
Gambar 4.3 Distribusi Kesalahan E2E Untuk Skenario 1	83
Gambar 4.4 <i>Confusion Matrix</i> Klasifikasi Biner HS dan NON-HS	88
Gambar 4.5 <i>Confusion Matrix</i> Tingkat Keparahan Tahap 3 Skenario 2.....	102
Gambar 4.6 Distribusi data benar E2E skenario 2	106
Gambar 4.7 <i>Boxplot</i> perbandingan skenario 1, 2, dan 3 (a) Tahap 1, (b) Tahap 2, (c) Tahap 2 <i>Hamming Loss</i> , (d) Tahap 3, (e) Akurasi sistem akhir antara 3 skenario	119
Gambar 4.8 <i>Boxplot</i> Pada Skenario 3	123
Gambar 4.9 Contoh <i>Error Analysis</i> secara keseluruhan	128
Gambar 4.10 Contoh <i>Error Analysis</i> Pada Tahap 2 Multi-label.....	130

DAFTAR TABEL

Tabel 2.1 Penelitian Terkait	9
Tabel 3.1 Atribut Pada Dataset	26
Tabel 3.2 Matriks <i>Co-Occurrence</i> Multi-label	32
Tabel 3.3 Proses <i>Text Cleaning</i>	35
Tabel 3.4 <i>Hyperparameter Fine-Tuning</i> IndoBERTweet.....	43
Tabel 3.5 Tabel Skenario Uji Coba.....	58
Tabel 4.1 Rekapitulasi Dataset Pada Setiap Tahap Skenario 1.....	61
Tabel 4.2 Rekapitulasi Hasil Percobaan Tahap 1 Pada Skenario 1.....	64
Tabel 4.3 Hasil Evaluasi Klasifikasi Biner Tahap 1 Pada Skenario 1	65
Tabel 4.4 Rekapitulasi Hasil Percobaan Tahap 2 Pada Skenario 1.....	67
Tabel 4.5 Metrik Evaluasi Klasifikasi Multi-label Tahap 2 Pada Skenario 1.....	68
Tabel 4.6 Distribusi Data Pada Klasifikasi Multi-label	70
Tabel 4.7 Rekapitulasi Jumlah Label Dalam 1 Tweet Pada Skenario 1	70
Tabel 4.8 <i>Confusion Matrix</i> Multi-label Tahap 2 Pada Skenario 1	71
Tabel 4.9 Rekapitulasi Hasil Percobaan Tahap 3 Pada Skenario 1.....	73
Tabel 4.10 Hasil Evaluasi Tingkat Keparahan Tahap 3 Pada Skenario 1.....	75
Tabel 4.11 Rekapitulasi Hasil Percobaan E2E Pada Skenario 1.....	80
Tabel 4.12 Hasil Distribusi Status E2E Pada Setiap Tahap Skenario 1.....	81
Tabel 4.13 Rekapitulasi Dataset Pada Setiap Tahap Skenario 2.....	84
Tabel 4.14 Rekapitulasi Hasil Percobaan Tahap 1 Pada Skenario 2.....	85
Tabel 4.15 Hasil Evaluasi Klasifikasi Biner Tahap 1 Pada Skenario 2	87
Tabel 4.16 Rekapitulasi Hasil Percobaan Tahap 2 Pada Skenario 2.....	89
Tabel 4.17 Distribusi Data Pada Klasifikasi Multi-label Pada Skenario 2	91
Tabel 4.18 Rekapitulasi Jumlah Label Dalam 1 Tweet Pada Skenario 2	93
Tabel 4.19 Hasil Evaluasi Klasifikasi Multi-Label (Tahap 2) Pada Skenario 2 ...	94
Tabel 4.20 <i>Confusion Matrix</i> Multi-label Pada Tahap 2 Skenario 2	95
Tabel 4.21 Rekapitulasi Hasil Percobaan Tahap 3 Pada Skenario 2.....	98
Tabel 4.22 Hasil Evaluasi Tingkat Keparahan Pada Tahap 3 Skenario 2.....	100
Tabel 4.23 Rekapitulasi Hasil Percobaan E2E Pada Skenario 2.....	103
Tabel 4.24 Hasil dan distribusi kesalahan pada evaluasi E2E skenario 2.....	105
Tabel 4.25 Rekapitulasi Hasil Evaluasi Pada Skenario 3	108
Tabel 4.26 Metrik Evaluasi Pendekatan Flat Klasifikasi	109
Tabel 4.27 <i>Confusion Matrix</i> Hasil Klasifikasi Flat Per Label.....	110
Tabel 4.28 Rekapitulasi Evaluasi Pada 3 Skenario	112
Tabel 4.29 Tabel <i>t-test</i> Antara Skenario 1, 2, dan 3.....	115

ABSTRAK

Maghfiroh, Siti Wasi'atul. 2026. **Klasifikasi Hierarkis Ujaran Kebencian Multi-Label Dengan IndoBERTweet Dan Regresi Ordinal**. 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. Irwan Budi Santoso, M.Kom.

Kata kunci: Ujaran kebencian, klasifikasi hierarkis, multi-label, IndoBERTweet, CORN, Twitter/X, NLP.

Ujaran kebencian pada media sosial merupakan permasalahan yang kompleks karena tidak hanya memerlukan deteksi ujaran kebencian, tetapi juga identifikasi kategori dan tingkat keparahannya. Penelitian ini bertujuan membangun sistem klasifikasi hierarkis multi-label berbasis IndoBERTweet untuk mendeteksi ujaran kebencian berbahasa Indonesia pada platform Twitter/X. Pendekatan hierarkis diterapkan melalui tiga tahap, yaitu klasifikasi biner, klasifikasi multi-label, dan klasifikasi ordinal menggunakan metode CORN untuk menentukan tingkat keparahan. Selain itu, penelitian ini menerapkan augmentasi data berbasis *Masked Language Modeling* (MLM) untuk mengatasi ketidakseimbangan data pada tahap multi-label. Hasil penelitian menunjukkan bahwa model klasifikasi biner memperoleh *Accuracy* sebesar 90,67% dan *F1-score* sebesar 90,37%. Pada tahap klasifikasi multi-label, model menghasilkan *Macro F1-score* sebesar 78,68% dan *Hamming Loss* sebesar 9,39%. Sementara itu, pada tahap klasifikasi ordinal diperoleh *Accuracy* sebesar 82,56% dan *Quadratic Weighted Kappa* sebesar 76,80%. Evaluasi *end-to-end system* menghasilkan performa sebesar 80,33%. Hasil uji *paired t-test* menunjukkan bahwa peningkatan performa akibat augmentasi data berbasis MLM belum signifikan secara statistik. Secara keseluruhan, metode klasifikasi hierarkis multi-label berbasis IndoBERTweet mampu mendeteksi ujaran kebencian dengan mempertimbangkan kategori multi-label dan tingkat keparahan secara ordinal pada data Twitter/X berbahasa Indonesia dengan performa yang baik.

ABSTRACT

Maghfiroh, Siti Wasi'atul. 2026. **Hierarchical Multi-Label Hate Speech Classification Using IndoBERTweet and Ordinal Regression**. Thesis. Informatics Engineering Study Program Faculty of Science and Technology Universitas Islam Negeri Maulana Malik Ibrahim Malang. Advisor: (I) Khadijah Fahmi Hayati Holle, M.Kom. (II) Dr. Irwan Budi Santoso, M.Kom.

Keywords: Hate speech, hierarchical classification, multi-label, IndoBERTweet, CORN, Twitter/X, NLP.

Hate speech detection on social media is a challenging task due to the need to identify both hate speech categories and severity levels. This study proposes a hierarchical multi-label classification system based on IndoBERTweet for Indonesian hate speech detection on Twitter/X. The system consists of three stages: binary classification, multi-label classification, and ordinal classification using the CORN method. To address data imbalance in the multi-label stage, Masked Language Modeling (MLM)-based data augmentation was applied. The results show that the binary classification model achieved an Accuracy of 90.67% and an F1-Score of 90.37%. The multi-label classification stage obtained a Macro F1-Score of 78.68% and a Hamming Loss of 9.39%, while the ordinal classification stage achieved an Accuracy of 82.56% and a Quadratic Weighted Kappa of 76.80%. The end-to-end evaluation reached 80.33%. Overall, the proposed hierarchical multi-label classification approach demonstrated good performance in detecting hate speech, identifying hate speech categories, and predicting severity levels in Indonesian Twitter/X data.

مستخلص البحث

مغفيرة، ستي واسعة. ٢٠٢٦. التصنيف الهرمي لخطاب الكراهية متعدد التصنيفات باستخدام IndoBERTweet والانحدار الترتيبي. بحث تخرج قسم الهندسة المعلوماتية، كلية العلوم والتكنولوجيا جامعة مولانا مالك إبراهيم الإسلامية الحكومية مالانج. المشرف الأول: خديجة فهمي حياتي هولي، الماجستير؛ المشرف الثاني الدكتور إروان بودي سانتوسو، الماجستير

الكلمات الرئيسية: خطاب الكراهية، التصنيف الهرمي، التصنيف متعدد التصنيفا، معالجة اللغة الطبيعية، IndoBertweet، معالجة اللغة الطبيعية، CORN، Twitter/X،

تُعَدُّ خطابات الكراهية على وسائل التواصل الاجتماعي مشكلةً معقدة، لأنها تتضمن تحديد فئات الخطاب ومستويات IndoBERTweet قائم على نموذج (Multi-Label) شدته. يهدف هذا البحث إلى بناء نظام تصنيف هرمي متعدد العلامات تم تطبيق النهج الهرمي من خلال ثلاث (Twitter/X) للكشف عن خطابات الكراهية باللغة الإندونيسية على منصة تويتر/إكس. لتحديد مستوى الشدة CORN مراحل، وهي التصنيف الثنائي، والتصنيف متعدد العلامات، والتصنيف الترتيبي باستخدام منهجية لمعالجة مشكلة Masked Language Modeling (MLM) كما اعتمد هذا البحث على تعزيز البيانات باستخدام تقنية %عدم توازن البيانات في مرحلة التصنيف متعدد العلامات. أظهرت نتائج البحث أن نموذج التصنيف الثنائي حقق دقة بلغت 90.67% بلغت Macro F1-Score 90.37%. وفي مرحلة التصنيف متعدد العلامات، حقق النموذج قيمة F1-Score وقيمة %بنسبة 9.39% أما في مرحلة التصنيف الترتيبي، فقد حقق النموذج دقة بلغت Hamming Loss 82.56 مع 78.68% أداءً بلغ (End-to-End) بنسبة 76.80%. كما أظهر التقييم الشامل للنظام Quadratic Weighted Kappa وقيمة ، أن التحسن الناتج عن تعزيز البيانات لم يكن ذا دلالة إحصائية. وبشكل عام paired t-test وأظهرت نتائج اختبار 80.33%، قدرةً جيدة على اكتشاف خطابات الكراهية IndoBERTweet أظهرت طريقة التصنيف الهرمي متعدد العلامات القائمة على والتعرف على الفئات متعددة العلامات، والتنبؤ بمستوى شدة خطابات الكراهية باللغة الإندونيسية.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Teknologi informasi telah mengalami perkembangan dan menjadikan media sosial sebagai sarana komunikasi utama dalam kehidupan masyarakat *modern*. Di Indonesia, jumlah pengguna media sosial terus meningkat setiap tahun, termasuk twitter/X, yang memiliki sekitar 24,7 juta pengguna aktif atau 8,9% dari populasi pada awal 2024 (Social, 2024). Twitter /X memungkinkan pengguna untuk berbagi informasi, mengekspresikan opini, dan membangun jejaring secara cepat dan luas (Syafuruddin *et al.*, 2024). Namun, selain manfaatnya Twitter/X juga menjadi sarana penyebaran ujaran kebencian (*hate speech*) yang menargetkan individu atau kelompok berdasarkan identitas agama, ras, etnis, maupun *gender*. Fenomena ini berpotensi menimbulkan ketegangan di dunia ruang digital sekaligus mengancam kohesi sosial masyarakat (Riadi & Fadlil, 2023).

Penyebaran komentar yang mengandung ujaran kebencian di Twitter/X berlangsung dengan intensitas tinggi, sehingga moderasi manual terbukti kurang efektif. Proses moderasi tradisional membutuhkan sumber daya besar, memakan waktu, dan sering terlambat dalam merespons konten bermasalah (Arifin & Mahdiana 2024). Kondisi ini menimbulkan kebutuhan mendesak terhadap sistem deteksi otomatis berbasis kecerdasan buatan dan *Natural Language Processing* yang mampu memproses *volume* data besar dengan cepat, akurat, dan efisien (Ibrohim & Budi, 2019).

Karakter ujaran kebencian cenderung kompleks, satu komentar dapat mengandung lebih dari satu kategori ujaran kebencian sekaligus. Pendekatan klasifikasi biner atau multi-label konvensional belum mampu menangkap kompleksitas hubungan antar kategori ujaran kebencian secara optimal (Ibrohim & Budi, 2019). Di sinilah pendekatan hierarki multi-label menjadi relevan. Dengan memanfaatkan struktur hierarkis, model dapat mengenali hubungan antar kategori, misalnya komentar yang bersifat diskriminatif terhadap agama sekaligus *gender*, sehingga prediksi menjadi lebih akurat dan kontekstual.

Aspek tingkat keparahan ujaran kebencian juga krusial. Banyak penelitian sebelumnya memperlakukan tingkat keparahan sebagai label diskrit, padahal tingkatan seperti *weak*, *moderate*, dan *strong* bersifat ordinal dan menentukan prioritas penanganan. Ujaran kebencian dengan tingkat *strong* memerlukan penanganan yang lebih cepat dibandingkan *weak* (Ibrohim & Budi, 2023). Mengabaikan sifat ordinal ini berisiko menyebabkan kesalahan penilaian dan *respons* yang tidak tepat. Dengan mempertimbangkan tingkat keparahan, sistem deteksi dapat memberikan rekomendasi tindakan yang lebih efektif, sehingga moderasi konten menjadi lebih terarah dan berdampak nyata (Sari *et al.*, 2022).

Karakter ujaran kebencian di media sosial memiliki beragam bentuk dan potensi dampak sosial, mulai dari sindiran hingga ujaran yang memicu konflik dan diskriminasi. Oleh karena itu, dalam moderasi konten digital diperlukan mekanisme yang mampu mengidentifikasi karakteristik ujaran secara komprehensif agar penanganan dapat dilakukan secara tepat dan efektif. Dalam ajaran Islam,

komunikasi yang baik dan buruk dibedakan secara tegas, sebagaimana ditegaskan dalam QS. Fussilat: 34.

وَلَا تَسْتَوِ الْحَسَنَةُ وَلَا السَّيِّئَةُ ۚ ادْفَعْ بِالَّتِي هِيَ أَحْسَنُ فَإِذَا الَّذِي بَيْنَكَ وَبَيْنَهُ عَدَاوَةٌ كَأَنَّهُ وَلِيٌّ حَمِيمٌ

“Tidaklah sama kebaikan dengan kejahatan. Tolaklah (kejahatan) dengan perilaku yang lebih baik sehingga orang yang ada permusuhan denganmu serta-merta menjadi seperti teman yang sangat setia.” (QS. Fussilat: 34)

Menurut Tafsir Ibnu Katsir, ayat ini menegaskan bahwa kebaikan dan keburukan tidak pernah setara, baik dalam perbuatan maupun ucapan. Allah memerintahkan agar keburukan yang diterima tidak dibalas dengan keburukan serupa, melainkan dengan sikap dan ucapan yang lebih baik, seperti kesabaran, kelembutan, dan akhlak yang terpuji. Dengan cara tersebut, permusuhan dapat berubah menjadi hubungan yang lebih baik. Ayat ini mengajarkan pengendalian diri dalam berinteraksi agar tidak terjerumus pada perbuatan atau ucapan yang dapat menimbulkan kerusakan sosial.

Dalam konteks moral dan etika, prinsip menjaga lisan sejalan dengan nilai-nilai Islam dalam QS. Al-Hujarat ayat 12:

يَا أَيُّهَا الَّذِينَ آمَنُوا اجْتَنِبُوا كَثِيرًا مِّنَ الظَّنِّ إِنَّ بَعْضَ الظَّنِّ إِثْمٌ وَلَا تَجَسَّسُوا وَلَا يَغْتَبِ بَعْضُكُم بَعْضًا أَيُحِبُّ أَحَدُكُمْ أَن يَأْكُلَ لَحْمَ أَخِيهِ مَيْتًا فَكَرِهْتُمُوهُ وَاتَّقُوا اللَّهَ إِنَّ اللَّهَ تَوَّابٌ رَّحِيمٌ ﴿١٢﴾

“Hai orang-orang yang beriman, jauhilah kebanyakan buruk sangka (kecurigaan), Karena sebagian dari buruk-sangka itu dosa. dan janganlah mencari-cari keburukan orang dan janganlah menggunjingkan satu sama lain. Adakah seorang diantara kamu yang suka memakan daging saudaranya yang sudah mati? Maka tentulah kamu merasa jijik kepadanya. dan bertakwalah kepada Allah. Sesungguhnya Allah Maha Penerima Taubat lagi Maha Penyayang.” (QS. Al-Hujarat :12)

Menurut Tafsir Ibnu Katsir, dijelaskan bahwa Allah Swt melarang hamba-hamba-Nya yang beriman dari banyak berprasangka buruk, yakni mencurigai keluarga dan kaum kerabat serta orang lain dengan tuduhan yang buruk yang bukan pada tempatnya. bahkan sebagian darinya termasuk dosa, karena itu, Allah memerintahkan agar hal tersebut di jauhi sepenuhnya sebagai bentuk kehati-hatian agar tidak menimbulkan keburukan yang lebih besar.(Haerudin *et al.*, 2025)

Berdasarkan penjelasan tafsir tersebut, jelas bahwa Islam menekankan kehati-hatian dalam berprasangka dan bertutur. Dalam konteks media sosial, nilai ini relevan dengan fenomena penyebaran ujaran kebencian yang seringkali muncul melalui komentar, unggahan, atau interaksi daring. Penelitian ini berupaya menghadirkan solusi berbasis teknologi melalui pengembangan sistem deteksi otomatis ujaran kebencian menggunakan model IndoBERTweet, sehingga konten yang berpotensi merugikan dapat diidentifikasi dan ditangani secara lebih efektif. Dengan demikian, penerapan teknologi tidak hanya berkontribusi pada moderasi konten, tetapi juga mendukung praktik komunikasi yang etis sesuai dengan nilai-nilai yang diajarkan dalam Islam.

Selain itu, penelitian ini diharapkan dapat mengembangkan model yang mampu menghasilkan klasifikasi yang lebih akurat dalam mengidentifikasi berbagai kategori ujaran kebencian beserta tingkat keparahannya. Dengan demikian, penelitian ini tidak hanya memberikan kontribusi bagi pengembangan kajian NLP, tetapi juga menawarkan solusi praktis untuk mendukung moderasi konten di media sosial serta mendorong terciptanya ruang lingkup digital yang lebih baik.

1.2 Rumusan Masalah

Bagaimana performa metode klasifikasi hierarkis multi-label dalam mendeteksi ujaran kebencian dengan mempertimbangkan kategori multi-label dan tingkat keparahan secara ordinal pada komentar Twitter/X berbahasa Indonesia, apabila dievaluasi menggunakan *F1-Score*, *Hamming loss*, *Macro F1-Score* dan *Quadratic Weighted Kappa*?

1.3 Batasan Masalah

Adapun batasan masalah dalam penelitian ini sebagai berikut.

1. Penelitian ini menggunakan dataset publik dari okkyibrohim (okkyibrohim, n.d.)
2. Fokus penelitian adalah deteksi ujaran kebencian dengan klasifikasi hierarkis multi-label berdasarkan kategori yang tersedia pada dataset.
3. Analisis tingkat keparahan ujaran kebencian dibatasi pada tiga level, yaitu *weak*, *moderate*, dan *strong*, yang diperlakukan sebagai variabel ordinal.
4. Evaluasi kinerja model dibatasi pada metrik *F1-Score*, *Hamming loss*, *Macro F1-Score* dan *Quadratic Weighted Kappa*.

1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk mengetahui performa metode klasifikasi hierarkis multi-label dalam mendeteksi ujaran kebencian pada komentar Twitter/X berbahasa Indonesia dengan kategori multi-label dan tingkat keparahan secara ordinal, yang dievaluasi menggunakan *F1-Score*, *Hamming loss*, *Macro F1-Score*, dan *Quadratic Weighted Kappa*.

1.5 Manfaat Penelitian

Dari uraian diatas, maka manfaat penelitian ini adalah sebagai berikut:

1. Penelitian ini diharapkan dapat membantu mengurangi dampak negatif ujaran kebencian dan mendorong terciptanya komunikasi yang lebih beretika di media sosial.
2. Mengembangkan riset di bidang NLP dan *machine learning*, khususnya terkait penerapan klasifikasi hierarkis multi-label dan regresi ordinal pada teks berbahasa indonesia.
3. Membantu moderasi konten di media sosial melalui deteksi ujaran kebencian yang lebih akurat, sehingga memberikan rasa aman bagi pengguna media sosial serta memudahkan pemilik akun atau pengelola *platform* dalam menjaga ruang digital yang kondusif.

BAB II

STUDI PUSTAKA

2.1 Penelitian Terkait

Menurut penelitian yang dilakukan (Ibrohim & Budi, 2019) meneliti tentang deteksi *hate speech* di *platform* Twitter/X menggunakan bahasa Indonesia, peneliti menggunakan metode multi-label dengan model SVM, *naive bayes* dan *random forest*. Penelitian ini menekankan bahwa satu tweet bisa memiliki lebih dari satu label, mencakup jenis ujaran, target, kategori, dan tingkat keparahan. Dengan hasil *random forest* dengan label *power-set* dan fitur *unigram* jadi yang terbaik dengan akurasi 76,16%. Akurasi naik jadi 77,36% saat ditambah fitur lain, tapi turun ke 66,12% jika juga memasukkan target, kategori, dan level *hate speech*.

Sedangkan menurut penelitian (Susanto *et al.*, 2024) membahas tentang bagaimana model *transformer* untuk klasifikasi ujaran kebencian berbahasa Indonesia. Penelitian ini menggunakan dataset IndoToxic2024 dengan pendekatan multi-label dan membandingkan performa beberapa varian model BERT. Hasil penelitian menunjukkan bahwa model IndoBERT memberikan performa terbaik dibandingkan model lain, dengan nilai *F1-Score* yang lebih tinggi dalam mendeteksi ujaran kebencian dan bahasa kasar.

Penelitian selanjutnya dilakukan oleh Bagestra (2024) membahas perbandingan metode klasik seperti SVM dan *naive bayes* dengan model *transformer* IndoBERT untuk deteksi ujaran kebencian multi-label di Twitter/X Indonesia. Hasil dari penelitian menunjukkan bahwa IndoBERT unggul dengan

akurasi 93% dan *F1-Score* 91%, sementara metode klasik jauh lebih rendah meskipun sudah ditingkatkan dengan data augmentasi.

Penelitian yang dilakukan oleh Prabowo (2019) membahas tentang identifikasi ujaran kebencian dan bahasa kasar di Twitter/X berbahasa Indonesia menggunakan pendekatan klasifikasi hierarkis multi-label. Penelitian ini menggunakan dataset yang telah dikategorikan ke dalam berbagai jenis ujaran kebencian. Metode yang diterapkan meliputi SVM, *random forest*, dan *naive bayes* dengan representasi fitur n-gram baik pada level kata maupun karakter. Hasil penelitian menunjukkan bahwa penerapan pendekatan hierarkis multi-label meningkatkan performa klasifikasi dibandingkan pendekatan multi-label biasa. Khususnya, kombinasi SVM dengan *word uni-gram* memberikan akurasi tertinggi sebesar 68,43%, menandakan efektivitas metode ini dalam mengenali berbagai kategori ujaran kebencian di media sosial.

Selanjutnya penelitian dilakukan oleh Hendrawan (2020) yang membahas klasifikasi multi-label ujaran kebencian dan bahasa kasar di Twitter/X Indonesia. Penelitian ini menerapkan metode *random forest decision tree* (RFDT), BiLSTM, dan BiLSTM dengan embedding BERT multilingual menggunakan transformasi label *binary relevance*, *label powerset*, dan *classifier chains*. Hasil penelitian menunjukkan bahwa kombinasi terbaik adalah RFDT dengan *classifier chains* tanpa *preprocessing* tambahan, yang mencapai akurasi 76,12%. Jika hanya dilihat dari label *hate speech*, RFDT menjadi model terbaik dengan akurasi 82,36%, diikuti BiLSTM (79,70%) dan BiLSTM+BERT (76,96%).

Berikut Tabel 2.1 dari penelitian terdahulu yang diuraikan pada paragraf sebelumnya.

Tabel 2.1 Penelitian Terkait

No	Nama	Judul	Dataset	Metode	Hasil
1.	Ibrohim & Budi, 2019	<i>Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter</i>	Dataset berasal dari 13.169 tweet berbahasa Indonesia (Ibrohim & Budi, 2018)	<i>Multi-label Text Classification</i> dengan algoritma (SVM, <i>naive bayes</i> , dan <i>random forest</i>) pakai transformasi (BR, LP, CC) serta fitur (<i>n-gram</i> , <i>orthography</i> , dan <i>lexicon</i>)	Kombinasi terbaik adalah <i>random forest decision tree</i> (RFDT) dengan Label <i>power-set</i> (LP) dan fitur <i>word unigram</i> yang mencapai akurasi 76,16%. Dengan tambahan fitur lain (<i>character quadgrams</i> , <i>sentiment</i> , dan <i>abusive lexicon</i>), akurasi meningkat menjadi 77,36%, namun turun ke 66,12% jika juga mencakup target, kategori, dan <i>level hate speech</i> .
2.	Susanto <i>et al.</i> , 2024	<i>IndoToxic2024: A Demographically-Enriched Dataset of Hate Speech and Toxicity Types for Indonesian Language</i>	Dataset berasal dari IndoToxic-2024	<i>fine-tuning</i> IndoBERTweet dengan <i>classification head</i> dan evaluasi menggunakan <i>F1-Score</i>	Model IndoBERTweet sebagai baseline memperoleh <i>macro F1</i> sebesar 0,78, menunjukkan dataset cukup menantang dan relevan untuk deteksi <i>hate speech</i> .
3.	Bagestra <i>et al.</i> , 2024	<i>Performance Assessment of Machine Learning and Transformer Models for</i>	Dataset berasal dari 13.169 tweet berbahasa Indonesia	SVM, <i>naive bayes</i> (TF-IDF) dengan <i>transformer</i>	IndoBERT memiliki performa tertinggi dengan akurasi

No	Nama	Judul	Dataset	Metode	Hasil
		<i>Indonesian Multi-Label Hate Speech Detection</i>	(Ibrohim & Budi, 2019)	IndoBERT (<i>fine-tuning</i>)	93% dan F1-Score 91%, mengungguli metode SVM dan <i>naive bayes</i>
4.	Prabowo <i>et al.</i> , 2019	<i>Hierarchical Multi-label Classification to Identify Hate Speech and Abusive Language on Indonesian Twitter.</i>	13.169 tweet berbahasa Indonesia (Ibrohim & Budi, 2019)	<i>Hierarchical Multi-label Classification</i> dengan Pendekatan (SVM, <i>random forest</i> , dan <i>naive bayes</i>) dengan fitur <i>n-gram (word & character)</i> .	metode <i>hierarchical multi-label classification</i> terbukti meningkatkan kinerja dibanding pendekatan biasa. Akurasi terbaik dicapai dengan SVM + <i>word uni-gram</i> = 68,43%.
5.	Hendrawan <i>et al.</i> , 2020	<i>Multilabel Classification of Hate Speech and Abusive Words on Indonesian Twitter Social Media.</i>	13.169 tweet berbahasa Indonesia (Ibrohim & Budi, 2019)	RFDT, BiLSTM, dan BERT <i>multilingual</i>	Model mBERT memiliki performa lebih baik dibandingkan metode RFDT dan BiLSTM dalam klasifikasi multi-label <i>hate speech</i> pada Twitter Indonesia dengan F1-Score terbaiknya 0.78

2.2 Ujaran kebencian di Media Sosial

Ujaran kebencian atau *Hate Speech* didefinisikan sebagai bentuk ekspresi yang menyebarkan, menghasut, mempromosikan, atau membenarkan kebencian terhadap individu maupun kelompok berdasarkan identitas tertentu seperti agama, ras, etnis, *gender*, atau orientasi seksual (Riadi & Fadlil, 2023). Di Indonesia, menjelaskan bahwa ujaran kebencian mencakup penghinaan, pencemaran nama

baik, penistaan, provokasi, serta penyebaran kebencian berbasis SARA. Twitter/X memiliki karakteristik unik yang memengaruhi bentuk ujaran kebencian, seperti keterbatasan panjang teks, penggunaan bahasa informal, singkatan, campuran bahasa, serta sarkasme (Ibrohim & Budi, 2019). Faktor ini membuat deteksi ujaran kebencian di Twitter/X lebih sulit dibanding teks formal.

Dampak ujaran kebencian tidak hanya merusak hubungan sosial dan memicu intoleransi, tetapi juga memengaruhi kondisi psikologis korban, seperti munculnya rasa tidak aman, menurunnya harga diri, stres, kecemasan, hingga depresi (Febryanti *et al.*, 2022). Dalam skala yang lebih luas, ujaran kebencian dapat menyebabkan polarisasi masyarakat dan konflik antar kelompok.

Dari sisi hukum, kebencian dapat berujung pada sanksi pidana termasuk dalam kategori tindak kejahatan yang diatur dalam peraturan perundang-undangan di Indonesia, khususnya dalam KUHP dan Undang-Undang Informasi dan Transaksi Elektronik (UU ITE) (Nuryani *et al.*, 2025). Ujaran kebencian tidak dapat dibenarkan sebagai bagian dari kebebasan berpendapat, melainkan merupakan perbuatan yang memiliki unsur melawan hukum, terutama jika mengandung kebencian berbasis suku, agama, ras, dan SARA (Rizky Pratama Putra Karo Karo, 2023). Selain itu, kebencian juga dikategorikan sebagai kejahatan yang diatur dalam UU No. 11 Tahun 2008 tentang ITE, sehingga memerlukan penegakan hukum yang tegas (G. Ambar Wulan, 2021).

2.3 IndoBERTweet

IndoBERTweet merupakan model bahasa berbasis arsitektur BERT yang dikembangkan khusus untuk teks berbahasa Indonesia di *platform* Twitter/X.

Model ini diperkenalkan oleh (Koto *et al.*, 2021) sebagai hasil adaptasi dari IndoBERT dengan penyesuaian kosakata dan pra-pelatihan tambahan menggunakan korpus besar tweet Indonesia. Tujuannya adalah agar model dapat memahami gaya bahasa informal, singkatan, dan penggunaan emoji yang umum di media sosial. Dengan kata lain, IndoBERTweet dirancang untuk menangkap konteks linguistik yang lebih sesuai dengan dinamika percakapan daring dibanding model bahasa umum seperti IndoBERT atau *Multilingual BERT* (Kusuma & Chowanda, 2023).

Secara umum, IndoBERTweet mempertahankan struktur *transformer* dari BERT-base namun memperluas daftar kosakatanya agar mencakup istilah dan token khas Twitter/X. Token baru tersebut diinisialisasi menggunakan rata-rata *embedding subword* BERT untuk mempercepat proses adaptasi (Hendrawan *et al.*, 2020). Pendekatan ini lebih efisien dibandingkan pra-pelatihan ulang dari awal, karena mampu menjaga stabilitas model sekaligus meningkatkan kinerja pada domain spesifik. Hasil pra-pelatihan menunjukkan bahwa IndoBERTweet mengalami peningkatan performa pada berbagai tugas pemrosesan bahasa alami berbasis media sosial, seperti analisis sentimen, deteksi ujaran kebencian, dan klasifikasi topik (Koto *et al.*, 2021).

Dalam konteks penelitian deteksi ujaran kebencian, IndoBERTweet sering digunakan sebagai model dasar (*backbone*) karena kemampuannya dalam menangkap karakteristik bahasa informal serta konteks sosial pada media Twitter/X. Model ini terbukti unggul dibanding IndoBERT pada dataset seperti IndoToxic2024 (Susanto *et al.*, 2024). Selain itu, arsitektur IndoBERTweet bersifat

fleksibel dan dapat diterapkan pada berbagai skema klasifikasi, termasuk klasifikasi biner, multi-label, maupun ordinal. Fleksibilitas ini memungkinkan pemodelan yang lebih komprehensif terhadap ujaran kebencian, mulai dari proses deteksi hingga pengukuran tingkat keparahan.

IndoBERTweet mendukung *fine-tuning* yang efisien melalui representasi token [CLS] yang digunakan sebagai representasi keseluruhan teks. Representasi ini kemudian dapat dimanfaatkan pada berbagai tugas klasifikasi, termasuk dalam skema multi-label maupun ordinal. Pengaturan parameter seperti *learning rate* yang rendah serta penerapan teknik *early stopping* dapat membantu menjaga kestabilan pelatihan, khususnya pada dataset media sosial yang memiliki tingkat noise tinggi (Wijanarko *et al.*, 2024). Dengan karakteristik tersebut, IndoBERTweet menjadi pilihan ideal untuk penelitian klasifikasi hierarkis ujaran kebencian karena mampu menggabungkan pemahaman kontekstual mendalam dengan ketahanan terhadap noise linguistik khas media sosial.

Pada penelitian ini, model yang digunakan adalah IndoBERTweet yang diadaptasi melalui proses *fine-tuning* pada dataset ujaran kebencian berbasis Twitter. Pemilihan model ini didasarkan pada kemampuannya dalam memahami karakteristik bahasa informal serta konteks sosial pada media sosial, sehingga sesuai untuk mendukung pendekatan klasifikasi hierarkis multi-label ujaran kebencian.

2.5 Klasifikasi Hierarkis

Klasifikasi hierarkis merupakan metode klasifikasi yang mempertimbangkan hubungan bertingkat antar label, di mana setiap label dapat

memiliki hubungan (*parent–child*). Pendekatan ini berbeda dengan klasifikasi datar (*flat classification*) yang memperlakukan semua label secara independen tanpa mempertimbangkan struktur hierarkis (Nedashkivskyi, 2024). Dalam metode ini, label disusun dari tingkat umum ke spesifik dengan mengikuti jalur tertentu dalam pohon atau graf hierarki. Pendekatan ini sangat relevan untuk deteksi ujaran kebencian, karena kategori ujaran kebencian biasanya memiliki struktur bertingkat (Susanto *et al.*, 2024). Sebagai contoh, label utama dapat berupa “*Hate Speech*”, yang kemudian bercabang menjadi subkategori berdasarkan target, seperti *religion*, etnis, atau *gender*, serta bentuk ujaran, seperti penghinaan, provokasi, atau diskriminasi (Ibrohim & Budi, 2019).

Beberapa strategi umum dalam klasifikasi hierarkis antara lain *Local Classifier per Node* (LCN) yaitu pendekatan di mana setiap *node* dalam hierarkis memiliki model tersendiri untuk memprediksi relevansi label, *Local Classifier per Parent Node* (LCPN) yaitu strategi yang melatih satu model pada setiap *parent node* untuk memprediksi label anak-anaknya, serta *Big-Bang Approach*, yaitu pendekatan yang memodelkan seluruh hierarkis sekaligus dalam satu proses pelatihan. Meskipun *Big-Bang Approach* dapat memanfaatkan informasi global dari seluruh struktur label, pendekatan ini memiliki kompleksitas yang tinggi, terutama ketika diterapkan pada hierarkis multi-label berskala besar (Nedashkivskyi, 2024).

2.4 Klasifikasi Biner

Klasifikasi biner merupakan pendekatan dasar dalam pembelajaran mesin yang bertujuan untuk mengelompokkan data ke dalam dua kelas yang saling

eksklusif (Kusuma & Chowanda, 2023). Setiap sampel dikategorikan sebagai anggota dari salah satu kelas, misalnya positif atau negatif, ya atau tidak, serta ujaran kebencian atau non-ujaran kebencian dalam konteks penelitian ini. Secara matematis, klasifikasi biner berfungsi untuk memetakan setiap input x ke salah satu label $y \in \{0,1\}$.

Model dilatih menggunakan data berlabel agar mampu memprediksi kelas baru berdasarkan pola yang telah dipelajari. Dalam deteksi ujaran kebencian, klasifikasi biner berperan untuk membedakan apakah suatu teks termasuk ujaran kebencian (*Hate Speech*) atau bukan (*Non-Hate Speech*). Tahap ini menjadi fondasi bagi proses klasifikasi bertingkat berikutnya, karena hasil identifikasi awal akan menentukan data yang diteruskan ke tahap klasifikasi multi-label (Susanto *et al.*, 2024).

2.5 Klasifikasi Multi-Label

Klasifikasi teks merupakan salah satu tugas utama dalam *Natural Language Processing* (NLP) yang bertujuan untuk mengkategorikan dokumen ke dalam kelas tertentu. Secara umum, terdapat 2 pendekatan dalam klasifikasi, yaitu klasifikasi *multi-class* dan klasifikasi multi-label. Pada klasifikasi *multi-class*, setiap data hanya memiliki satu label yang dapat dipilih atau output bersifat tunggal. Misalnya sebuah teks hanya dapat dikategorikan sebagai “positif” atau “negatif”. Sedangkan klasifikasi multi-label memungkinkan satu data memiliki lebih dari satu label aktif secara bersamaan, sehingga satu teks dapat termasuk kedalam beberapa kategori sekaligus (Bagestra *et al.*, 2024).

Dalam konteks deteksi ujaran kebencian, pendekatan multi-label menjadi lebih relevan karena satu komentar pada media sosial, seperti Twitter/X dapat mengandung lebih dari satu jenis kebencian, misalnya terkait agama, ras, dan gender secara bersamaan (Hendrawan *et al.*, 2020). Oleh karena itu, penggunaan *multi-label classification* memungkinkan model untuk merepresentasikan kompleksitas makna dalam teks secara lebih akurat dibandingkan dengan pendekatan *multiclass*.

Dalam menyelesaikan permasalahan multi-label, terdapat beberapa pendekatan umum seperti *Binary Relevance (BR)*, *Classifier Chains (CC)*, dan *Label Powerset (LP)* yang digunakan untuk mentransformasikan label ke dalam bentuk yang dapat diproses oleh model klasifikasi (Ibrohim & Budi, 2023). Namun, penelitian ini tidak secara khusus berfokus pada perbandingan metode transformasi label tersebut, melainkan menggunakan pendekatan langsung berbasis model *deep learning* untuk menangani klasifikasi multi-label.

Dengan demikian, penggunaan klasifikasi multi-label dalam penelitian ini bertujuan untuk merepresentasikan kompleksitas data ujaran kebencian secara lebih akurat, sehingga model dapat mengidentifikasi berbagai kategori kebencian dalam satu instance teks secara bersamaan. Dalam implementasinya, penelitian ini tidak berfokus pada perbandingan metode transformasi label, melainkan menggunakan pendekatan langsung berbasis model *deep learning* untuk menangani klasifikasi multi-label.

2.6 Tingkat Keparahannya Menggunakan Regresi Ordinal

Regresi ordinal merupakan pendekatan statistik yang digunakan untuk memodelkan hubungan antara satu atau lebih variabel independen dengan variabel dependen yang bersifat kategorikal berurutan (*ordered categorical variable*) (Lu et al., 2022). Berbeda dengan klasifikasi biasa yang memperlakukan setiap kelas secara independen, regresi ordinal mempertahankan urutan logis antar kelas, seperti rendah, sedang, tinggi (Shaw et al., 2025).

Beberapa penelitian terbaru menunjukkan bahwa ujaran kebencian sebaiknya tidak diklasifikasikan hanya secara biner (ada/tidak ada), melainkan juga perlu diperhatikan tingkat keparahannya untuk menilai seberapa kuat dampak ujaran tersebut terhadap korban atau komunitas (Ayele et al., 2024). Penelitian *Measuring Hate Speech Corpus* oleh (Sachdeva et al., 2022) menggunakan anotasi berurutan untuk mengukur *hate intensity* menggunakan teori *Rasch Measurement*, yang menunjukkan bahwa *severity* dapat dimodelkan sebagai variabel ordinal kontinu. Sementara itu, studi oleh (Ayele et al., 2024) juga menyoroti pentingnya mengukur intensitas ujaran kebencian (*hate intensity*) guna membedakan ujaran yang bersifat ringan, ofensif, hingga ekstrem.

Untuk bahasa Indonesia, penelitian *IndoBERT-Based Ensemble Learning for Multi-Level Multi-Label Hate Speech Detection* (Rokhim et al., 2024) dengan dataset IndoToxic2024 menunjukkan adanya kebutuhan untuk model yang mampu mengenali *multi-level severity* secara terstruktur. Hal ini memperkuat relevansi penggunaan regresi ordinal dalam sistem klasifikasi hierarkis, di mana tingkat

keparahan menjadi tahap akhir setelah identifikasi ujaran kebencian dan kategorinya.

Dalam penelitian ini, pendekatan regresi ordinal digunakan untuk memodelkan tingkat keparahan ujaran kebencian yang terdiri dari tiga level, yaitu *weak*, *moderate*, dan *strong*. Ketiga level tersebut memiliki hubungan berurutan, sehingga tidak tepat jika diperlakukan sebagai kelas yang independen. Dengan menggunakan regresi ordinal, model dapat mempertimbangkan hubungan antar level tersebut, sehingga kesalahan prediksi antar kelas yang berdekatan (misalnya antara *weak* dan *moderate*) tidak dianggap seburuk kesalahan antara *weak* dan *strong*.

2.6.1 Konsep Dasar Regresi Ordinal

Regresi ordinal merupakan pengembangan dari *logistic regression* yang digunakan untuk memodelkan data kategorikal yang memiliki urutan (*ordered categorical*). Model ini menghitung probabilitas kumulatif dari suatu observasi termasuk dalam atau di bawah kategori tertentu menggunakan fungsi *logit*, seperti ditunjukkan pada persamaan 2.1 berikut.

$$\text{logit}[P(Y \leq j)] = \alpha_j - \beta^T X \quad (2.1)$$

Keterangan:

Y : variabel target ordinal (tingkat keparahan)

j : kategori ke- j dalam urutan (1, 2, ..., k)

α_j : ambang batas (*threshold*) pada kategori ke- j

β : vektor parameter regresi

X : vektor fitur input

Persamaan 2.1 menunjukkan bahwa model memprediksi probabilitas kumulatif suatu data berada pada atau di bawah kategori ke- j dengan mempertimbangkan

pengaruh variabel input X terhadap tingkat keparahan. Untuk mendukung hal tersebut, digunakan fungsi logit yang merupakan transformasi dari probabilitas ke dalam bentuk *log-odds*, sebagaimana ditunjukkan pada Persamaan 2.2.

$$\text{logit}(P) = \log\left(\frac{P}{1-P}\right) \quad (2.2)$$

Pada Persamaan 2.2, fungsi logit mengubah nilai probabilitas P ke dalam bentuk *log-odds*, yaitu perbandingan antara peluang suatu kejadian terjadi dengan peluang tidak terjadinya. Transformasi ini memungkinkan nilai probabilitas yang awalnya berada pada rentang 0 hingga 1 menjadi tidak terbatas, sehingga dapat dimodelkan menggunakan pendekatan linear dalam regresi ordinal.

Berdasarkan formulasi pada Persamaan 2.1 model regresi ordinal ini dikenal sebagai *Proportional Odds Model* (POM). Model ini memodelkan probabilitas kumulatif bahwa variabel target berada pada atau di bawah kategori tertentu. Dengan pendekatan ini, hubungan berurutan antar kelas dapat dipertahankan, sehingga lebih sesuai untuk kasus yang memiliki tingkatan, seperti level keparahan ujaran kebencian (Lu *et al.*, 2022).

Dalam penelitian ini, variabel X merepresentasikan fitur berupa *embedding* teks yang dihasilkan oleh model IndoBERTweet. *Embedding* tersebut berupa vektor numerik yang merepresentasikan makna semantik dari teks dan digunakan sebagai input dalam model.

Meskipun *Proportional Odds Model* (POM) banyak digunakan dalam pendekatan statistik tradisional, model ini memiliki keterbatasan ketika diterapkan pada arsitektur *deep learning*, khususnya dalam integrasi dengan representasi

embedding dari model *transformer*. Oleh karena itu, diperlukan pendekatan yang lebih fleksibel dan kompatibel dengan *neural network*, seperti *Conditional Ordinal Regression for Neural Networks* (CORN).

2.6.2 *Conditional Ordinal Regression for Neural Networks*

Dalam pengembangan model berbasis *deep learning*, regresi ordinal menjadi pendekatan yang relevan ketika variabel target memiliki urutan tingkat, seperti tingkat keparahan ujaran kebencian (*severity level*). Salah satu metode yang secara khusus dikembangkan untuk menangani masalah tersebut adalah *Conditional Ordinal Regression for Neural Networks* (CORN) yang diperkenalkan oleh (Xintong Shi *et al.*, 2023). Regresi ordinal memperlakukan kelas sebagai kategori berurutan yang memiliki hubungan tingkat.

Metode CORN mengubah permasalahan regresi ordinal menjadi serangkaian submasalah biner yang bersifat kondisional (*conditional binary tasks*). Jika terdapat K kelas berurutan, CORN membentuk $K - 1$ fungsi biner dengan target probabilitas yang dinyatakan pada persamaan 2.3

$$P(Y > r_k | X) = \sigma(z_k) \quad (2.3)$$

Keterangan:

r_k : ambang batas (*threshold*) ke- k dalam urutan kelas

z_k : output linear dari model untuk fungsi biner ke- k

σ : fungsi sigmoid

Pada persamaan 2.3 setiap fungsi biner memprediksi apakah kelas aktual Y lebih besar dari level r_k tertentu. Model kemudian mempelajari probabilitas kumulatif tersebut secara kondisional untuk menjaga konsistensi urutan prediksi (*rank-consistency*). Prediksi akhir ditentukan dengan menghitung jumlah keputusan

benar secara bertingkat menggunakan fungsi indikator, seperti ditunjukkan pada persamaan 2.4

$$\hat{y} = 1 + \sum_{k=1}^{K-1} I(P(Y > r_k) > 0.5) \quad (2.4)$$

Keterangan:

$I(\cdot)$: fungsi indikator yang bernilai 1 jika kondisi benar dan 0 jika salah.

$\hat{y}$: prediksi kelas ordinal

K : jumlah kelas ordinal

Persamaan 2.4 digunakan untuk menentukan kelas prediksi akhir dengan menjumlahkan jumlah kondisi yang terpenuhi, yaitu ketika probabilitas $P(Y > r_k | X)$ melebihi ambang batas tertentu. Hasil penjumlahan tersebut kemudian ditambah satu untuk memperoleh kelas ordinal akhir.

Selama proses pelatihan, setiap submasalah biner dilatih menggunakan *Binary Cross-Entropy Loss (BCE)* untuk meminimalkan kesalahan prediksi, sebagaimana dirumuskan pada 2.5 berikut.

$$\mathcal{L}_{\text{CORN}} = \frac{1}{K-1} \sum_{r=1}^{K-1} \text{BCE}(P(Y > r | X), t_r) \quad (2.5)$$

Keterangan:

$\mathcal{L}_{\text{CORN}}$: fungsi loss pada metode CORN

t_r : label target biner untuk submasalah ke- r

BCE : *Binary Cross-Entropy*

Persamaan 2.5 digunakan sebagai fungsi loss dalam metode CORN, yang menghitung rata-rata kesalahan dari setiap submasalah biner menggunakan *Binary Cross Entropy (BCE)*. Setiap submasalah membandingkan probabilitas prediksi dengan label target, sehingga model dapat belajar memprediksi urutan kelas secara konsisten. Di mana $t_r=1$ jika $Y > r$, dan 0 jika sebaliknya.

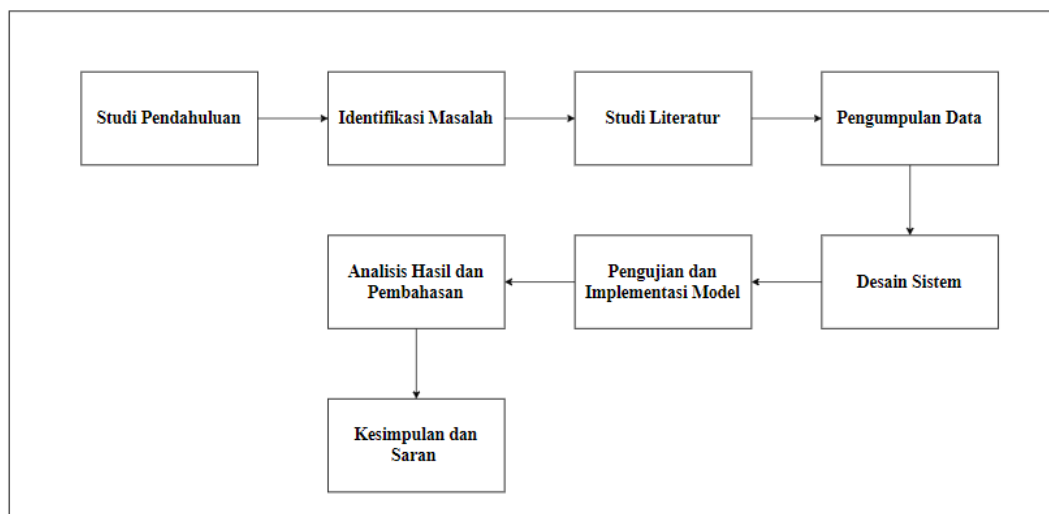
Dalam penelitian ini, metode CORN diimplementasikan dengan memanfaatkan *embedding* teks yang dihasilkan oleh IndoBERTweet sebagai input model. *Embedding* tersebut kemudian diproses melalui *fully connected layer* untuk menghasilkan output pada setiap fungsi biner dalam CORN. Output ini merepresentasikan probabilitas bahwa suatu data berada di atas level tertentu, seperti lebih dari *weak* atau lebih dari *moderate*. Hasil dari seluruh fungsi biner tersebut kemudian digabungkan untuk menentukan level keparahan akhir (*weak*, *moderate*, atau *strong*), sehingga model tidak hanya mampu mengklasifikasikan ujaran kebencian, tetapi juga mempertahankan hubungan urutan antar tingkat keparahan secara konsisten.

BAB III

METODE PENELITIAN

3.1 Desain Penelitian

Desain penelitian dibuat untuk memberi arah yang jelas dalam pelaksanaan penelitian, mulai dari melakukan studi literatur sampai dengan penarikan kesimpulan dan saran. Dengan adanya rancangan ini diharapkan penelitian ini dapat dilakukan secara lebih teratur dan sistematis, sehingga hasil yang dicapai sesuai dengan tujuan yang ingin dicapai. Desain penelitian dapat dilihat pada gambar 3.1 dibawah ini.



Gambar 3.1 Desain Penelitian

Berdasarkan desain Gambar 3.1 tersebut tahap pertama yaitu melakukan studi pendahuluan untuk mengetahui dan memahami permasalahan apa yang akan diteliti, setelah itu dilakukan identifikasi permasalahan yang muncul dengan tujuan untuk merumuskan masalah secara jelas dan spesifik, sehingga arah penelitian menjadi fokus dan terarah. Tahap selanjutnya yaitu melakukan studi literatur dari

penelitian terdahulu dengan mengumpulkan teori, konsep, serta hasil penelitian terdahulu yang relevan dengan topik penelitian, studi ini berfungsi sebagai dasar ilmiah dalam merancang model atau sistem yang akan dibangun serta sebagai landasan dalam menentukan metode penelitian. Dilanjutkan dengan

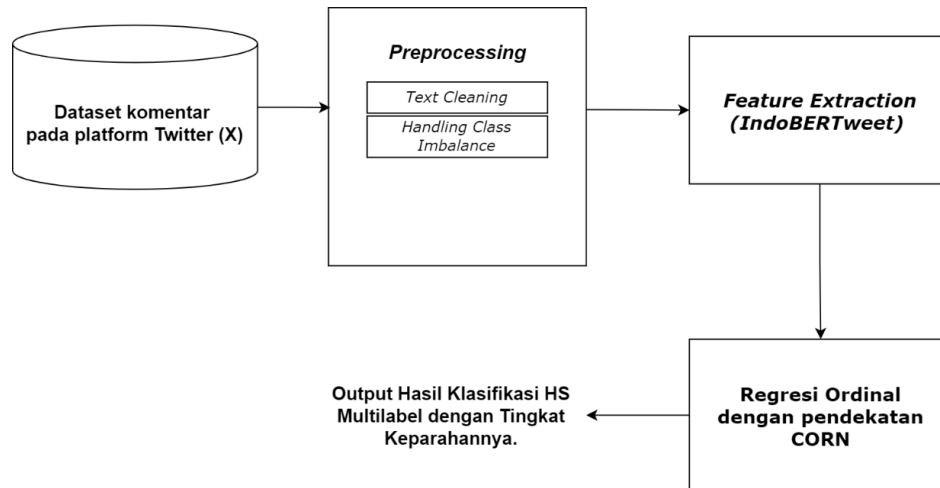
Tahap selanjutnya yaitu pengumpulan data, dimana penelitian ini menggunakan dataset publik dari (okkyibrohim, n.d.) berisi tentang tweet ujaran kebencian yang tersedia pada *platform github*. Berdasarkan dataset yang telah diperoleh, peneliti merancang sistem klasifikasi berbasis model IndoBERTweet. Rancangan sistem ini mencakup arsitektur klasifikasi hierarkis yang terdiri dari tiga tahap : klasifikasi biner (HS atau NON-HS), klasifikasi kategori multi-label, dan analisis tingkat keparahan menggunakan regresi ordinal.

Pada tahap selanjutnya Pada tahap ini dilakukan *training*, *testing*, dan *validation* terhadap model IndoBERTweet yang telah diimplementasikan sesuai desain. Pengujian dilakukan untuk mengukur performa model pada tiap tahap klasifikasi menggunakan metrik evaluasi seperti *F1-Score*, *Hamming Loss*, *Macro F1-Score* dan *Quadratic Weighted Kappa*. Hasil dari pengujian tersebut kemudian dianalisis untuk melihat sejauh mana model mampu mengidentifikasi ujaran kebencian dengan baik. Dan pada tahap terakhir berisi kesimpulan dari hasil penelitian, termasuk efektivitas pendekatan klasifikasi hierarkis multi-label dan regresi ordinal dalam mendeteksi ujaran kebencian berbahasa indonesia.

3.2 Desain Sistem

Penelitian ini menggunakan pendekatan klasifikasi hierarkis multi-label dan regresi ordinal untuk mengetahui tingkat keparahan dari sebuah tweet menjadi

beberapa tahap hierarkis. Desain sistem yang diterapkan dalam riset ini ditunjukkan melalui alur diagram pada Gambar 3.2.



Gambar 3.2 Desain Sistem

Desain sistem pada Gambar 3.2 diawali dengan input dan *Exploratoring Dataset Analysis* (EDA) pada dataset tweet berbahasa Indonesia. Selanjutnya dilakukan tahap preprocessing berupa *text cleaning* dengan menghapus elemen yang tidak relevan, seperti URL, *username*, *hashtag*, emoji, dan simbol tertentu, serta melakukan *case folding* untuk mengubah seluruh huruf menjadi huruf kecil.

Penelitian ini tidak menerapkan *stemming* maupun normalisasi karakter secara agresif karena model IndoBERTweet telah dilatih menggunakan data media sosial berbahasa Indonesia dengan karakteristik bahasa informal, sehingga bentuk kata asli dipertahankan agar representasi token tetap sesuai dengan model *pre-trained* yang digunakan (Bagestra *et al.*, 2024). proses selanjutnya yaitu *handling class imbalance* untuk mengatasi data yang tidak *balance* antara kelas minoritas dan mayoritas.

Setelah melakukan *preprocessing* dilakukan *split data* menjadi 3 yaitu *training set*, *validation set*, *testing set*. Untuk penanganan *class imbalance* dilakukan pada data *training* menggunakan *text* augmentasi. Pada *split data* ini dilakukan hanya pada tahap 1. Selanjutnya dilakukan *feature extraction* menggunakan *transformer* dengan pendekatan IndoBERTweet yang menghasilkan vektor representatif pada setiap tahap hierarkis. Proses selanjutnya yaitu, pemodelan menggunakan klasifikasi hierarkis dalam 3 tahap yaitu klasifikasi biner, klasifikasi multi-label, dan tingkat keparahan menggunakan regresi ordinal dengan pendekatan (CORN). Langkah terakhir dilanjutkan dengan evaluasi model dan menghasilkan klasifikasi HS hierarkis multi-label dengan menggunakan regresi ordinal.

3.3 Dataset

Peneliti disini menggunakan dataset publik dari (okkyibrohim, n.d.) berasal dari *platform* Github, yang berjumlah 13.169 data dengan total 13 kolom pada dataset tersebut. Untuk atribut pada dataset ini ditampilkan pada Tabel 3.1

Tabel 3.1 Atribut Pada Dataset

Kolom	Deskripsi	Tipe Data
Tweet	Berisi komentar dari Twitter/X	<i>String</i>
HS	Ujaran yang mengandung kebencian terhadap seseorang atau kelompok.	<i>Integer</i>
<i>Abusive</i>	Kata-kata kasar atau hinaan ke orang lain	<i>Integer</i>
<i>HS_Individual</i>	Ujaran kebencian yang ditujukan kepada individu tertentu	<i>Integer</i>
<i>HS_Group</i>	Ujaran kebencian yang ditujukan kepada suatu kelompok	<i>Integer</i>
<i>HS_Religion</i>	Ujaran kebencian yang menyerang agama atau kepercayaan.	<i>Integer</i>
<i>HS_Race</i>	Ujaran kebencian yang menyerang ras atau etnis	<i>Integer</i>
<i>HS_Physical</i>	Ujaran kebencian yang menyerang kondisi atau penampilan fisik.	<i>Integer</i>
<i>HS_Gender</i>	Ujaran kebencian yang menyerang gender atau identitas gender.	<i>Integer</i>
<i>HS_Other</i>	Ujaran kebencian dalam kategori lain yang tidak termasuk dalam kategori sebelumnya (misalnya aspek sosial atau umum).	<i>Integer</i>
<i>HS_Weak</i>	Ujaran kebencian tingkat ringan	<i>Integer</i>
<i>HS_Moderate</i>	Ujaran kebencian tingkat sedang	<i>Integer</i>

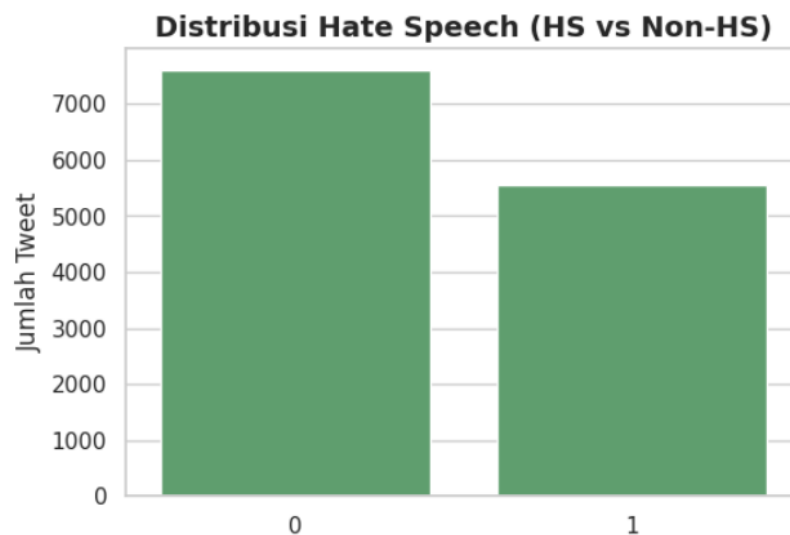
Kolom	Deskripsi	Tipe Data
HS_Strong	Ujaran kebencian tingkat berat	<i>Integer</i>

Pada Tabel 3.1 menunjukkan dataset berisi komentar dari twitter/X untuk mendeteksi ujaran kebencian dan kata kasar. Setiap kolom bertipe *integer* berisi nilai 0 atau 1 yang menunjukkan ada atau tidaknya kategori tertentu dalam komentar. Kategori tersebut meliputi target (individu atau kelompok), jenis serangan, dan tingkat keparahan ujaran kebencian (ringan, sedang, berat). Berikut Gambar 3.3 yang memperlihatkan isi dari dataset tersebut.

	Tweet	HS	Abusive	HS_Individual	HS_Group	HS_Religion	HS_Race	HS_Physical	HS_Gender	HS_Other	HS_Weak	HS_Moderate	HS_Strong
0	- disaat semua cowok berusaha melacak perhatia...	1	1	1	0	0	0	0	0	1	1	0	0
1	RT USER: USER siapa yang telat ngasih tau elu?...	0	1	0	0	0	0	0	0	0	0	0	0
2	41. Kadang aku berfikir, kenapa aku tetap perc...	0	0	0	0	0	0	0	0	0	0	0	0
3	USER USER AKU ITU AKU\n\nKU TAU MATAMU SIPIT T...	0	0	0	0	0	0	0	0	0	0	0	0
4	USER USER Kaum cebong kapir udah kelihatan dong...	1	1	0	1	1	0	0	0	0	0	1	0
5	USER Ya bani taplak dkk \xf0\x9f\x98\x84\xf0\x...	1	1	0	1	0	0	0	0	1	0	1	0
6	deklarasi pilkada 2018 aman dan anti hoax warg...	0	0	0	0	0	0	0	0	0	0	0	0
7	Gue baru aja kelar re-watch Aldnoah Zero!!! pa...	0	1	0	0	0	0	0	0	0	0	0	0
8	Nah admin belanja satu lagi port terbaik nak m...	0	0	0	0	0	0	0	0	0	0	0	0
9	USER Enak lg klo smbil ngewe'	0	1	0	0	0	0	0	0	0	0	0	0
10	Setidaknya gw punya jari tengah buat lu, sebel...	1	1	1	0	0	0	0	0	1	1	0	0
11	USER USER USER USER BANCI KALENG MALU GA BISA ...	1	1	1	0	0	0	0	1	0	1	0	0
12	Kalo belajar ekonomi mestinya jago memprivatis...	0	0	0	0	0	0	0	0	0	0	0	0
13	Aktor huruhara 98 Prabowo S Ingin lengserkan p...	1	0	1	0	0	0	0	0	1	1	0	0
14	USER Bu guru enakan jadi jablay atau guru esde...	1	1	1	0	0	0	0	1	0	1	0	0
15	USER USER USER USER USER USER Lawan bicara gw ...	1	1	1	0	1	0	0	0	0	1	0	0
16	Belakangan ini kok fikiran ampas banget ya'	0	0	0	0	0	0	0	0	0	0	0	0
17	Ari sarua beki mah repeh monyet\xf0\x9f\x98\x8...	0	1	0	0	0	0	0	0	0	0	0	0
18	Jadi cowo itu harus Gentle kalo ga Gentle itu ...	1	1	0	1	0	0	0	1	0	0	1	0
19	USER Slga mnr bom \xf0\x9f\x98\x82'	0	0	0	0	0	0	0	0	0	0	0	0
20	Asw ya tapi gua jarang ngambek, tacut wkkwkwkw...	0	1	0	0	0	0	0	0	0	0	0	0
21	USER kalo kamu noob pasti peluang disakitin ny...	0	0	0	0	0	0	0	0	0	0	0	0
22	USER Joko Widodo dinilai sebagai presiden terl...	0	0	0	0	0	0	0	0	0	0	0	0

Gambar 3.3 Isi Dataset

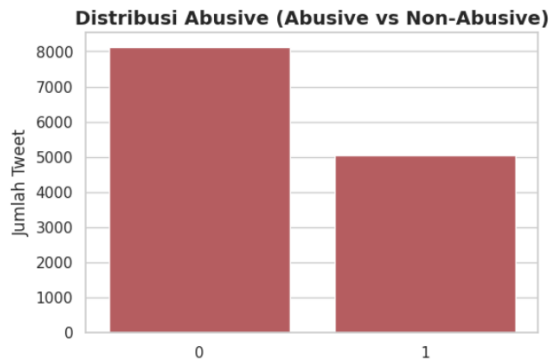
Pada Gambar 3.3 menampilkan *sample* dari dataset yang akan digunakan pada penelitian ini. Kolom tweet merupakan kolom utama yang berisi komentar bahasa indonesia. Setelah itu terdapat kolom *hate speech* dan *Non hate speech*, kolom tersebut digunakan untuk indikator apakah tweet mengandung ujaran kebencian atau tidak. Berikut pada Gambar 3.4 ditampilkan visualisasi distribusi data pada kolom HS.



Gambar 3. 4 Visualisasi EDA HS dan Non-HS

Pada Gambar 3.4 tersebut, kolom HS digunakan sebagai indikator untuk menunjukkan apakah suatu tweet mengandung ujaran kebencian atau tidak. Visualisasi ini memperlihatkan perbandingan antara data yang termasuk HS dan NON-HS.

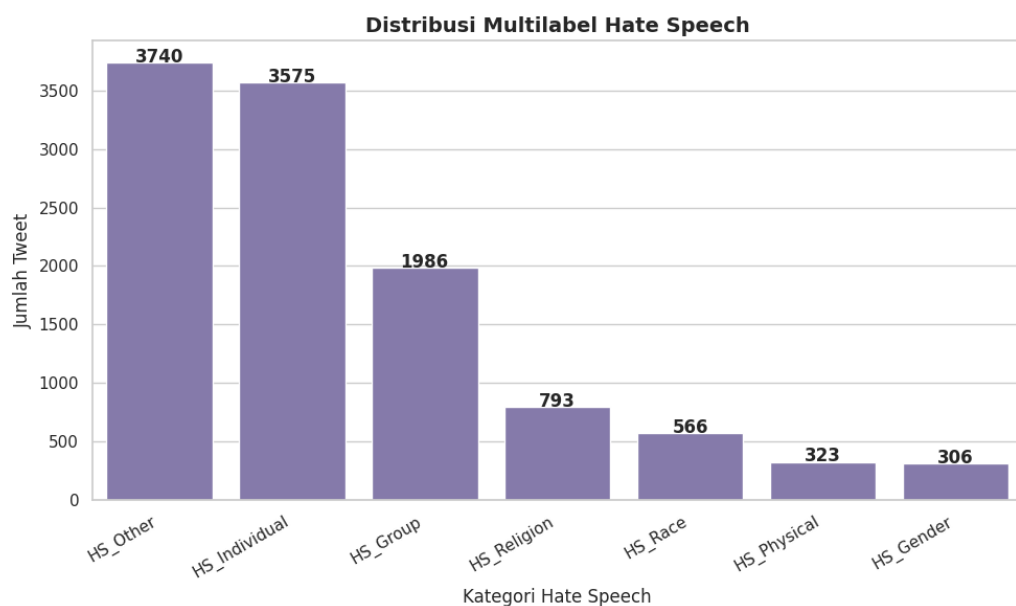
Didalam dataset juga terdapat kolom *abusive* sebagai informasi pendukung. Kolom tersebut digunakan untuk indikator apakah tweet mengandung kata kasar/*abusive language* atau tidak, berikut visualisasi persebaran data untuk kolom *abusive* bisa dilihat pada Gambar 3.5



Gambar 3. 5 Visualisasi EDA *Abusive* dan *Non Abusive*

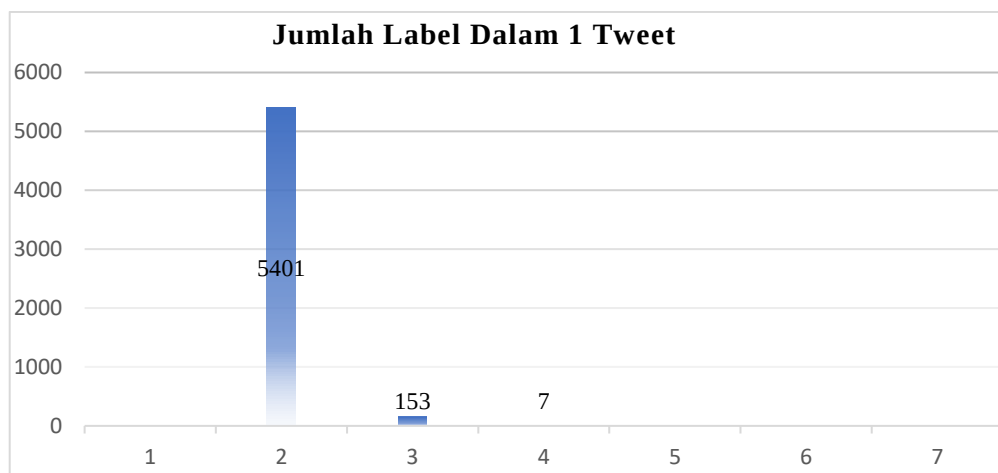
Berdasarkan Gambar 3.5, terlihat perbandingan antara *tweet* yang mengandung kata kasar (*abusive*) dan yang tidak. Visualisasi ini menunjukkan distribusi data pada kolom *Abusive*.

Selanjutnya, dilakukan analisis distribusi data pada kolom kategori ujaran kebencian. Kolom ini digunakan pada tahap klasifikasi multi-label, di mana setiap ujaran kebencian dibagi berdasarkan jenis dan sasarannya. Data terdiri dari 7 label, berikut persebaran datanya dapat dilihat pada Gambar 3.6.



Gambar 3.6 Visualisasi EDA *Multi-label Hate Speech*

Pada Gambar 3.6 dapat dilihat bahwa persebaran data pada setiap label tidak seimbang. *HS_Other* memiliki jumlah paling banyak dan *HS_Gender* memiliki jumlah paling sedikit. Pada dataset ini 1 tweet bisa memiliki lebih dari 1 label, berikut pada Gambar 3.7 yang menampilkan jumlah label dalam 1 tweet



Gambar 3.7 Jumlah Label Dalam 1 Tweet

Pada Gambar 3.7 menampilkan ketidakseimbangan data yang dimiliki setiap label, tweet yang memiliki paling banyak datanya dengan label berjumlah 2, dan label yang memiliki 3 label hanya berjumlah 153 dan yang paling banyak labelnya namun sedikit tweetnya berjumlah 4 sebanyak 7. Dan tidak ada tweet yang memiliki label sebanyak 1, 5, 6, dan 7. Sehingga ketidakseimbangan data akan mempengaruhi bagaimana proses menangani pada tahap klasifikasi.

Untuk menganalisis hubungan antar label dalam data ujaran kebencian, digunakan matriks *co-occurrence* yang menunjukkan frekuensi kemunculan dua label secara bersamaan dalam satu tweet. Analisis ini bertujuan untuk mengidentifikasi pola keterkaitan antar kategori ujaran kebencian serta melihat apakah terdapat kecenderungan label tertentu muncul secara bersamaan. Hasil

matriks ini dapat memberikan gambaran mengenai kompleksitas data yang bersifat multi-label.

Tabel 3.2 Matriks *Co-Occurrence* Multi-label

label	HS_ <i>Religion</i>	HS_ <i>Race</i>	HS_ <i>Individual</i>	HS_ <i>Group</i>	HS_ <i>Physical</i>	HS_ <i>Gender</i>	HS_ <i>Other</i>
HS_ <i>Religion</i>	759	92	257	502	18	29	0
HS_ <i>Race</i>	92	548	148	400	1	5	0
HS_ <i>Individual</i>	257	148	3460	0	277	241	2604
HS_ <i>Group</i>	502	400	0	1924	40	59	1018
HS_ <i>Physical</i>	18	1	277	40	317	24	0
HS_ <i>Gender</i>	29	5	241	59	24	300	0
HS_ <i>Other</i>	0	0	2604	1018	0	0	3622

Berdasarkan matriks *co-occurrence* yang ditunjukkan pada Tabel 3.2, terlihat bahwa label *HS_Individual* memiliki jumlah kemunculan tertinggi yaitu sebesar 3460 data, diikuti oleh *HS_Other* sebesar 3622 data dan *HS_Group* sebesar 1924 data. Hal ini menunjukkan bahwa ujaran kebencian dalam dataset didominasi oleh serangan terhadap individu.

Untuk hubungan antar label, kombinasi yang paling dominan adalah *HS_Individual* dengan *HS_Other* sebesar 2604 data, serta *HS_Group* dengan *HS_Other* sebesar 1018 data, yang menunjukkan bahwa label “Other” sering muncul sebagai label tambahan dalam satu tweet. Selain itu, kombinasi *HS_Religion* dengan *HS_Group* sebesar 502 data dan *HS_Race* dengan *HS_Group* sebesar 400 data mengindikasikan bahwa ujaran kebencian berbasis identitas cenderung diarahkan kepada kelompok.

Di sisi lain, tidak ditemukan kemunculan bersamaan antara *HS_Individual* dan *HS_Group* (nilai 0), yang menunjukkan bahwa kedua label tersebut bersifat eksklusif dan merepresentasikan target yang berbeda. Secara keseluruhan, hasil ini

menegaskan bahwa label dalam data tidak berdiri sendiri, melainkan saling berkaitan, sehingga pendekatan *multi-label classification* lebih tepat digunakan.

Analisis data selanjutnya yaitu untuk kolom tingkat keparahan. Kolom ini dibagi menjadi 3 tingkatan yang masing-masing menunjukkan seberapa kuat atau berbahayanya ujaran. Klasifikasi tingkat ujaran kebencian dalam penelitian ini mengacu pada pembagian yang digunakan oleh (Ibrohim & Budi, 2019) dalam kajian deteksi ujaran kebencian di media sosial Indonesia. Pembagian ini bersifat teknis dan digunakan untuk menggambarkan karakteristik serta potensi dampak sosial dari ujaran kebencian, bukan sebagai penilaian moral atau keagamaan.

1. HS_Weak: ujaran kebencian ringan

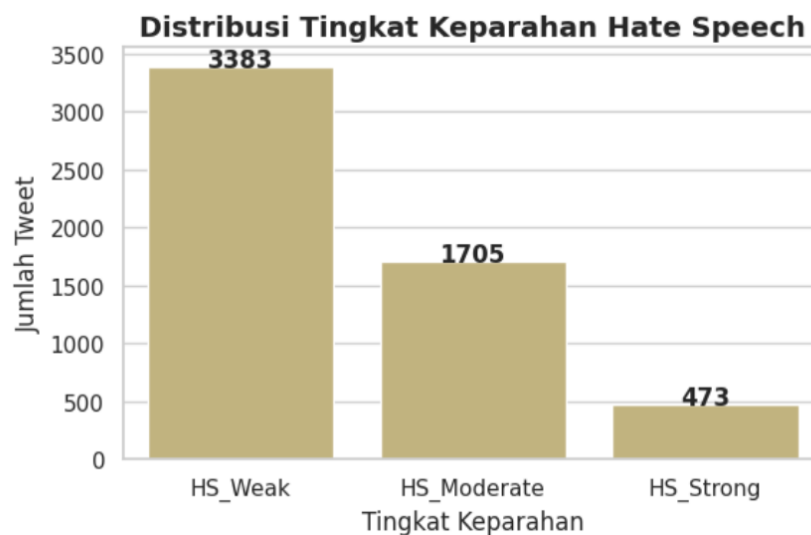
Ujaran kebencian kategori lemah merupakan ujaran berupa hinaan, makian, atau fitnah yang ditujukan kepada individu secara personal tanpa mengandung ajakan atau provokasi untuk menimbulkan konflik terbuka. Ujaran ini umumnya bersifat personal dan memiliki dampak sosial yang terbatas, sehingga dalam praktik moderasi konten tidak selalu menjadi prioritas utama penanganan.

2. HS_Moderate: ujaran kebencian sedang

Ujaran kebencian kategori sedang mencakup ujaran berupa hinaan, penistaan, stereotip, atau pelabelan yang ditujukan kepada kelompok tertentu tanpa disertai ajakan atau provokasi konflik secara langsung. Ujaran ini berpotensi menimbulkan ketegangan antar kelompok, terutama di ruang media sosial, namun dampaknya masih terbatas pada konflik di ranah digital.

3. HS_Strong: ujaran kebencian berat.

Ujaran kebencian kategori kuat mencakup ujaran yang ditujukan kepada individu maupun kelompok dan mengandung unsur ajakan atau provokasi untuk menimbulkan konflik terbuka. Ujaran ini memiliki potensi dampak sosial yang luas dan serius, serta dapat memicu konflik di dunia nyata, sehingga perlu diprioritaskan dalam proses deteksi dan penanganan. Berikut visualisasi persebaran datanya pada Gambar 3.8



Gambar 3.8 Visualisasi EDA Tingkat Keparahan (*Severity Level*)

Pada Gambar 3.8 menunjukkan bahwa *HS_Weak* memiliki jumlah yang lebih banyak daripada *HS_Moderate*. *HS_Strong* memiliki jumlah yang minoritas sehingga bisa dikatakan persebaran data pada tingkat keparah belum seimbang.

3.4 Preprocessing

Setelah mendapatkan dataset, sistem dilanjutkan dengan tahap *preprocessing* data yang berfungsi untuk mempersiapkan teks agar siap diproses pada tahap analisis selanjutnya.

3.4.1 Text Cleaning

Text cleaning merupakan tahap awal dalam preprocessing teks yang bertujuan untuk membersihkan data agar lebih rapi, konsisten, dan siap diproses pada tahap selanjutnya. Pada penelitian ini, proses *text cleaning* dilakukan dengan menghapus beberapa elemen yang tidak relevan, seperti URL/link, *username*, *hashtag*, *retweet marker*, emoji, angka yang tidak diperlukan, *punctuation* berlebih, *whitespace* ganda, serta karakter *non-ASCII* atau simbol tertentu yang dapat mengganggu proses representasi teks. Selain itu, dilakukan juga proses *case folding*, yaitu mengubah seluruh huruf menjadi huruf kecil agar penulisan teks lebih konsisten. Penelitian ini tidak menerapkan *stemming* maupun normalisasi karakter secara agresif karena model IndoBERTweet telah dilatih menggunakan data media sosial berbahasa Indonesia dengan karakteristik bahasa informal. Oleh karena itu, bentuk kata asli dipertahankan agar konteks dan representasi token tetap sesuai dengan model *pre-trained* yang digunakan. Berikut Tabel 3.3 yang berisi perbandingan data sebelum dan sesudah *text cleaning*.

Tabel 3.3 Proses *Text Cleaning*

Sebelum <i>Text Cleaning</i>	Sesudah <i>Text Cleaning</i>
Jadi cowo itu harus Gantle kalo ga Gantle itu namanya BANCIII !'	jadi cowo itu harus gantle kalo ga gantle itu namanya banciii

3.4.2 Handling Class Imbalance

Class imbalance merupakan kondisi ketika jumlah data pada setiap kelas tidak seimbang (Henning *et al.*, 2023), misalnya jumlah data ujaran kebencian (*hate speech*) lebih sedikit dibandingkan data bukan ujaran kebencian (*non-hate speech*). Kondisi ini dapat menyebabkan model menjadi bias, yaitu lebih cenderung memprediksi kelas mayoritas dan mengabaikan kelas minoritas, sehingga

menurunkan performa dalam mengenali ujaran kebencian. Untuk mengatasi hal tersebut, penelitian ini menggunakan teknik *oversampling* berbasis data teks menggunakan *text augmentation*. Pendekatan ini dilakukan dengan memperbanyak data pada kelas minoritas dengan cara menghasilkan variasi teks baru yang tetap mempertahankan makna semantik dari data asli. Pada penelitian ini, *text augmentation* dilakukan menggunakan pendekatan *Masked Language Modeling* (MLM), di mana beberapa token pada teks minoritas dimask dan kemudian diprediksi ulang oleh model bahasa untuk menghasilkan kalimat baru yang kontekstual. Teknik ini memungkinkan terbentuknya data sintetis yang lebih natural dan bervariasi tanpa mengubah label kelas.

Pseudocode 3.1 Algorithm MLM Text Augmentation

```

Input: Dataset D, minority class C_min, masking ratio p, target
size N_target
Output: Augmented dataset D'

1. Pisahkan data kelas minoritas:
   D_min ← {x | (x, y) ∈ D dan y = C_min}
2. Inisialisasi dataset hasil:
   D_aug ← D
3. While |D_min| < N_target do
   a. Pilih satu teks x dari D_min
   b. Tokenisasi teks x menjadi token T = {t1, t2, ..., tL}
   c. Hitung jumlah token yang akan dimask:
      n_mask ← floor (p × L)
   d. Pilih n_mask token secara acak dari T
   e. Ganti token terpilih dengan token [MASK]
   f. Gunakan model MLM untuk memprediksi token pengganti
   g. Bentuk teks baru x_aug dari hasil prediksi
   h. Tambahkan (x_aug, C_min) ke D_aug
   i. Tambahkan x_aug ke D_min
   End While
4. Return D_aug

```

Algoritma ini bekerja dengan memilih data dari kelas minoritas, kemudian menghasilkan variasi teks baru melalui proses masking dan prediksi token menggunakan model MLM. Proses ini diulang hingga jumlah data kelas minoritas

mencapai target yang diinginkan atau seimbang dengan kelas mayoritas. Dengan cara ini, data sintetis yang dihasilkan tetap mempertahankan konteks dan makna semantik dari teks asli.

3.5 Split Data

Sebelum proses pelatihan dilakukan, dataset dibagi menjadi tiga bagian, yaitu data latih (*training set*) sebesar 70%, data validasi (*validation set*) sebesar 15%, dan data uji (*testing set*) sebesar 15%. Pembagian ini bertujuan agar model dapat dilatih, divalidasi, dan diuji secara terstruktur. Data latih digunakan untuk menyesuaikan parameter model selama proses pelatihan. Data validasi digunakan untuk memantau performa model sementara itu, data uji berfungsi untuk mengevaluasi kemampuan generalisasi model terhadap data baru yang tidak digunakan pada tahap pelatihan maupun validasi.

3.6 Feature Extraction

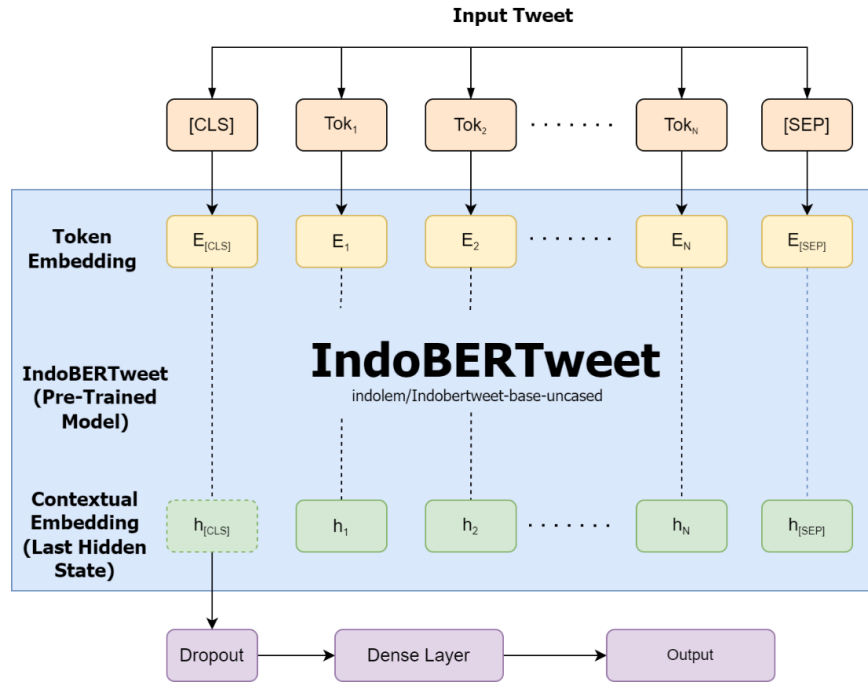
Setelah melalui tahap *preprocessing*, teks yang telah dibersihkan perlu diubah menjadi bentuk numerik agar dapat diproses oleh model. Tahap ini disebut *feature extraction* atau *text representation*. Arsitektur model yang digunakan pada penelitian ini adalah IndoBERTweet, yaitu versi BERT yang telah dilatih secara khusus menggunakan data Twitter berbahasa Indonesia. Hal ini dilakukan agar model dapat memahami karakteristik teks informal seperti singkatan, *emoticon*, dan gaya bahasa khas media sosial yang sering muncul pada dataset tweet.

Setelah melalui tahap *preprocessing*, teks yang telah dibersihkan perlu diubah ke dalam representasi numerik agar dapat diproses oleh model. Pada

penelitian ini, proses representasi teks dilakukan menggunakan model *indolem/indoBERTweet-base-uncased*, yaitu model berbasis *transformer* yang telah dilatih menggunakan data Twitter berbahasa Indonesia. Penggunaan IndoBERTweet bertujuan agar model dapat memahami karakteristik bahasa informal pada media sosial, seperti singkatan, bahasa tidak baku, dan gaya penulisan khas tweet.

Pada tahap ini, teks diproses menggunakan *AutoTokenizer* dari *library transformers* dengan model *indolem/indobertweet-base-uncased* untuk mengubah teks menjadi token numerik yang sesuai dengan *vocabulary model pre-trained*. Hasil tokenisasi kemudian direpresentasikan dalam bentuk *embedding* sehingga dapat diproses oleh arsitektur *transformer*.

Selanjutnya, digunakan *BertForSequenceClassification* yang telah disesuaikan dengan IndoBERTweet untuk membangun model klasifikasi multi-label dan regresi ordinal. Model ini dilatih (*fine-tuning*) pada dataset khusus *hate speech* untuk mengklasifikasikan tweet ke dalam kategori ujaran kebencian dan level keparahannya. Proses *fine-tuning* memungkinkan model mempelajari pola spesifik pada dataset sehingga prediksi menjadi lebih akurat dibandingkan menggunakan model pra-latih tanpa penyesuaian. Berikut arsitektur proses ekstraksi fitur menggunakan IndoBERT ditunjukkan pada Gambar 3.9.



Gambar 3. 9 Arsitektur *Fine-Tuning* IndoBERTweet

Dari Gambar 3.9 langkah-langkah utama pada tahap ini mencakup tokenisasi, embedding, dan encoding menggunakan model IndoBERTweet. Teks mentah $T = [w_1, w_2, \dots, w_n]$ diubah menjadi token menggunakan *tokenizer* bawaan *IndoBERTweet*, menghasilkan token [CLS] untuk representasi keseluruhan kalimat dan token [SEP] sebagai penanda akhir kalimat, dengan rumus pada 3.2

$$X = \text{Tokenizer}(T) = [x_{\text{CLS}}, x_1, x_2, \dots, x_n, x_{\text{SEP}}] \quad (3.2)$$

Token [CLS] digunakan sebagai representasi global kalimat yang akan menjadi input ke lapisan klasifikasi. Token [SEP] menandai akhir kalimat, dan setiap token kata x_i akan diproses lebih lanjut ke tahap embedding.

Setiap token yang dihasilkan dari tahap tokenisasi kemudian diubah menjadi representasi vektor melalui proses *embedding*. *Embedding* ini merupakan hasil

penjumlahan dari tiga komponen utama, yaitu *token embedding*, *positional embedding*, dan *segment embedding*, sebagaimana ditunjukkan pada rumus 3.3

$$e_i = E_{\text{token}}(x_i) + E_{\text{pos}}(i) + E_{\text{seg}}(x_i) \quad (3.3)$$

Keterangan:

e_i : vektor embedding token ke-i

E_{token} : embedding dari token xi

E_{pos} : embedding posisi token ke-i

E_{seg} : embedding segmen

m : jumlah token

d : dimensi embedding

Setelah itu, seluruh *embedding* token membentuk matriks representasi kalimat, dengan persamaan 3.4

$$E = [e_{\text{CLS}}, e_1, \dots, e_n, e_{\text{SEP}}] \quad (3.4)$$

Dimana $E \in R^{(n+2) \times d}$ dan $d = 768$ yang merupakan dimensi embedding untuk IndoBERTweet. Setiap komponen *embedding* berperan dalam memberikan informasi semantik kata, posisi dalam urutan kalimat, dan segmen teks, sehingga model mampu memahami konteks kata dalam urutan yang berbeda.

Selanjutnya matriks *embedding* tersebut kemudian diproses melalui lapisan *Transformer encoder* IndoBERTweet yang terdiri dari beberapa lapisan *multi-head self-attention* dan *feed-forward network*. Proses encoding pada lapisan ke- l dapat dinyatakan dengan persamaan 3.5 sebagai berikut:

$$\begin{aligned} E &= H^{(l)} = \text{TransformerLayer}^{(l)}(H^{(l-1)}), \quad l = 1, \dots, L \\ H &= H^{(L)} = [h_{\text{CLS}}, h_1, \dots, h_n, h_{\text{SEP}}] \end{aligned} \quad (3.5)$$

Hasil akhir dari proses *encoding* berupa representasi kontekstual untuk setiap token. Outputnya dari setiap token h_1 merepresentasikan kata beserta

konteksnya dalam kalimat. Token [CLS] h_{CLS} menyimpan representasi global kalimat yang akan digunakan untuk klasifikasi.

Berikut contoh perhitungannya menggunakan teks hasil *preprocessing* diatas yaitu: ['jadi', 'cowok', 'harus', 'gentle', 'kalo', 'tidak', 'gentle', 'nama', 'banci']. Selanjutnya dilakukan tokenisasi oleh *tokenizer* bawaan model IndoBERTweet (dari *library transformers*) menggunakan persamaan 3.2. *Tokenizer* ini secara otomatis menambahkan token khusus [CLS] di awal dan [SEP] di akhir kalimat sebagai penanda awal dan akhir urutan teks. Hasil tokenisasi ditunjukkan sebagai berikut: ['[CLS]', 'jadi', 'cowok', 'harus', 'gentle', 'kalo', 'tidak', 'gentle', 'nama', 'banci', '[SEP]']. Token [CLS] digunakan sebagai representasi keseluruhan kalimat untuk keperluan klasifikasi, sedangkan [SEP] menandai batas akhir dari teks. Setiap token kemudian dikonversi ke bentuk numerik berdasarkan kamus (*vocabulary*) IndoBERTweet sebelum memasuki tahap *embedding*.

Setelah proses tokenisasi, setiap token diubah menjadi representasi numerik melalui tahapan *embedding*. IndoBERTweet menghasilkan vektor berdimensi 768 untuk setiap token, yang diperoleh dari penjumlahan tiga komponen utama, yaitu *token embedding*, *positional embedding*, dan *segment embedding*.

Setelah pembentukan *embedding*, setiap token direpresentasikan dalam bentuk vektor numerik yang akan mekanisme *self-attention* untuk memahami hubungan antar-token dalam satu kalimat. *Self-attention* memungkinkan model memperhatikan seluruh kata dalam kalimat secara bersamaan untuk menentukan kata mana yang paling relevan terhadap konteks tertentu. Setiap token diubah menjadi tiga representasi, yaitu *Query (Q)*, *Key (K)*, dan *Value (V)*, yang diperoleh

melalui perkalian antara vektor *embedding* dengan matriks bobot hasil pelatihan model. Nilai perhatian antar-token kemudian dihitung menggunakan rumus *Scaled Dot-Product Attention* pada 3.6

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3.6)$$

Keterangan

Q : representasi *query* token yang sedang dievaluasi

K : representasi key token lain dalam kalimat

V : representasi *value* yang membawa informasi semantik

d_k : dimensi vektor kunci (digunakan untuk normalisasi)

setiap token dalam kalimat akan menghitung tingkat keterkaitannya dengan token lain melalui nilai perhatian (α_{ij}). Nilai ini diperoleh dari hasil operasi antara *query* (Q_i) milik token yang sedang dianalisis dan *key* (K_j) dari seluruh token lain dalam kalimat. Perhitungan bobot perhatian dilakukan menggunakan fungsi *softmax* terhadap hasil perkalian titik ($Q_i K_j^T$) yang telah dinormalisasi dengan akar dimensi vektor kunci ($\sqrt{d_k}$). Persamaannya ditunjukkan pada Persamaan (3.7)

$$\alpha_{ij} = \text{softmax}\left(\frac{Q_i K_j^T}{\sqrt{d_k}}\right) \quad (3.7)$$

Persamaan ini menghasilkan distribusi bobot yang menggambarkan seberapa besar pengaruh setiap token terhadap token lain dalam satu kalimat. Semakin tinggi nilai α_{ij} , semakin besar kontribusi token ke- j terhadap pemaknaan token ke- i .

Selanjutnya, nilai perhatian tersebut digunakan untuk menghitung representasi kontekstual baru dari setiap token (h_i) dengan cara mengalikan bobot perhatian dengan *value* (V_j) dari masing-masing token. Proses ini dirumuskan pada Persamaan 3.8:

$$h_i = \sum_j \alpha_{ij} V_j \quad (3.8)$$

Hasil dari perhitungan ini berupa vektor kontekstual h_i yang mengandung makna kata berdasarkan keseluruhan konteks kalimat. Dengan demikian, setiap kata tidak hanya merepresentasikan dirinya sendiri, tetapi juga membawa informasi dari kata-kata lain yang memiliki keterkaitan makna.

Selanjutnya tahap *Fine-Tuning* IndoBERTweet, model dilatih untuk menyesuaikan bobot dan parameter internalnya berdasarkan data pelatihan yang telah diproses. Proses ini menggunakan algoritma AdamW, yaitu pengembangan dari *Adam optimizer* yang menambahkan komponen *weight decay* untuk mencegah *overfitting*. Tujuan utamanya adalah membuat model tetap sederhana namun mampu belajar pola data dengan stabil. AdamW mempercepat konvergensi selama pelatihan dengan memanfaatkan dua momentum, yaitu momentum pertama dan momentum kedua. Agar model dapat bekerja secara optimal, proses *fine tuning* dilakukan dengan parameter Tabel 3.4:

Tabel 3.4 *Hyperparameter Fine-Tuning* IndoBERTweet

<i>parameter</i>	Nilai
<i>Tokenizer</i>	<i>AutoTokenizer</i> IndoBERTweet
<i>Maximum Sequence Length</i>	512
<i>Batch Size</i>	16
<i>Learning Rate</i>	5e-5
<i>Optimizer</i>	AdamW
β_1	0.9
β_2	0.999
ϵ (epsilon)	1e-8
<i>Weight Decay</i> (λ)	0.01
<i>Epoch</i>	5
<i>Scheduler</i>	<i>Linear Learning Rate Scheduler</i>
<i>Random Seed</i>	42

Berdasarkan Tabel 3.4, *training* model dilakukan menggunakan *optimizer* AdamW dengan *learning rate* sebesar $5e-5$ dan *weight decay* 0.01 untuk mengurangi *overfitting*. Model dilatih selama 5 *epoch* dengan *batch size* 16 dan maksimum panjang token 512. Konfigurasi yang sama digunakan pada seluruh eksperimen, baik pendekatan flat maupun hierarkis, untuk memastikan perbandingan performa yang adil

Momentum pertama m_t merupakan rata-rata eksponensial dari gradien, sedangkan momentum kedua v_t adalah rata-rata eksponensial dari kuadrat gradien yang dihitung selama *backpropagation*. Misalnya pada iterasi pertama ($t=1$) nilai momentum pertama dan kedua diatur menjadi 0 karena belum ada perhitungan sebelumnya. Proses pembaruan kemudian dilakukan menggunakan gradien pertama yang telah diperoleh. Pembaharuan pertama dan momentum ke dua dihitung dengan rumus pada 3.9 dan 3.10:

$$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t \quad (3.9)$$

$$v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2 \quad (3.10)$$

Keterangan

m_t : estimasi momentum pertama pada langkah t

v_t : estimasi momentum kedua pada langkah t

β_1 : faktor *decay* momentum pertama

g_t : gradien pada langkah t

β_2 : faktor *decay* momentum kedua

Gradien g_t merupakan perubahan bobot pada iterasi ke- t , sedangkan β_1 dan β_2 adalah faktor *decay* untuk momentum pertama dan kedua. Nilai m_t dan v_t diperbarui di setiap iterasi agar model dapat belajar lebih efisien dengan memanfaatkan informasi dari gradien sebelumnya.

Koreksi bias kemudian dilakukan untuk menyesuaikan estimasi momentum pada awal pelatihan, karena nilai m_t dan v_t pada iterasi awal cenderung kecil dan belum stabil.

Koreksi ini dihitung menggunakan rumus pada 3.11 dan 3.12

$$\widehat{m}_t = \frac{m_t}{1 - \beta_1^t} \quad (3.11)$$

$$\widehat{v}_t = \frac{v_t}{1 - \beta_2^t} \quad (3.12)$$

Keterangan

$\widehat{m}_t$ = momentum pertama setelah koreksi bias

$\widehat{v}_t$ = momentum kedua setelah koreksi bias

t = iterasi (langkah ke- t)

Dengan penerapan ketiga langkah tersebut, algoritma AdamW dapat melakukan pembaruan parameter secara adaptif dan stabil. Komponen *weight decay* membantu mengontrol kompleksitas model dengan mengurangi nilai bobot besar secara bertahap, sehingga mencegah *overfitting* selama proses *fine-tuning* IndoBERTweet, dimana parameter model pada iterasi 0 adalah $\theta_0 = 1$, maka pembaruan parameter dihitung menggunakan rumus pada 3.13

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \epsilon} - \eta \cdot \lambda \cdot \theta_t \quad (3.13)$$

Keterangan

θ_{t+1} : parameter model yang diperbarui

θ_t : parameter model sebelumnya

η : *learning rate*

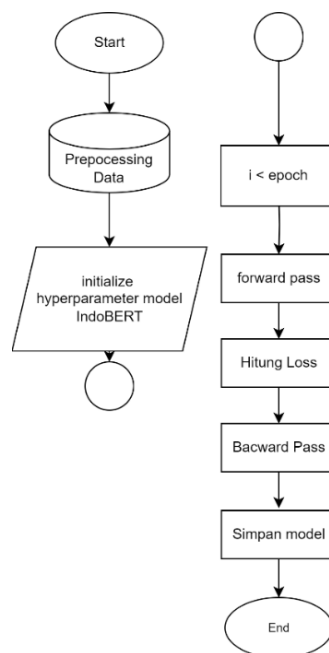
ϵ : konstanta kecil mencegah pembagian dengan nol

λ : *weight decay*

Dalam penerapan model IndoBERTweet, algoritma *AdamW* membantu proses pelatihan agar parameter model dapat diperbarui secara efisien dan stabil. Metode ini juga mencegah *overfitting*, sehingga model mampu menghasilkan performa yang baik dalam mengklasifikasikan ujaran kebencian pada teks media sosial.

3.7 Pemodelan

Selanjutnya tahap pemodelan dalam penelitian ini melibatkan tiga proses utama klasifikasi hierarkis yang meliputi klasifikasi utama, klasifikasi kategori multi-label, dan klasifikasi tingkat keparahan. Masing-masing tahap memiliki mekanisme pelatihan dan pengujian tersendiri sesuai dengan karakteristik tugas klasifikasinya. Secara umum, proses pelatihan dilakukan dengan menggunakan optimisasi AdamW, *learning rate* 5e-5, *batch size* 16, dan mekanisme *early stopping* untuk menjaga performa model. Proses pelatihan model IndoBERTweet secara umum digambarkan pada Gambar 3.12 berikut.



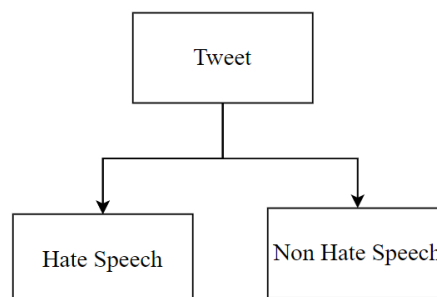
Gambar 3.10 Flowchart Proses Pelatihan Model IndoBERTweet

Pada Gambar 3.12 *flowchart* tersebut menggambarkan proses pelatihan model yang dimulai dari tahap *preprocessing* data teks, kemudian dilanjutkan dengan inisialisasi model beserta *hyperparameter* yang digunakan. Selanjutnya,

model dilatih melalui beberapa iterasi (*epoch*), di mana pada setiap iterasi dilakukan *forward pass* untuk menghasilkan prediksi, perhitungan *loss* untuk mengukur kesalahan, serta *backward pass* untuk memperbarui bobot model. Proses ini berlangsung berulang hingga jumlah epoch terpenuhi, setelah itu model yang telah dilatih disimpan sebagai hasil akhir untuk digunakan pada tahap pengujian atau prediksi.

3.7.1 Tahap 1: Klasifikasi Biner

Tahap pertama bertujuan untuk mengidentifikasi apakah suatu teks termasuk ke dalam ujaran kebencian (*Hate Speech*) atau bukan (*Non-Hate Speech*). Proses ini merupakan langkah awal dalam pendekatan hierarkis, di mana model dilatih untuk melakukan klasifikasi biner terhadap data masukan. Model akan mempelajari pola linguistik dan semantik dari teks berdasarkan representasi vektor hasil *embedding* IndoBERTweet. Jika hasil prediksi menunjukkan bahwa teks termasuk kategori *Hate Speech*, maka data tersebut akan diteruskan ke tahap berikutnya untuk klasifikasi kategori yang lebih spesifik. Sebaliknya, jika teks termasuk *Non-Hate Speech*, maka proses klasifikasi berhenti pada tahap ini. Seperti alur klasifikasi pada Gambar 3.13



Gambar 3. 11 Alur klasifikasi biner

Input pada tahap ini yaitu vektor [CLS] dari IndoBertweet h_{CLS} . Dengan rumus aktivasi sigmoid untuk memprediksi biner pada (3.14)

$$\hat{y} = \sigma(W_1 h_{CLS} + b_1) \quad (3.14)$$

Keterangan:

$\hat{y}$: probabilitas prediksi ujaran kebencian

b_1 : bias *layer*

W_1 : bobot *layer fully connected* untuk klasifikasi biner

σ : fungsi aktivasi sigmoid

Dengan perhitungan fungsi *Loss* pada persamaan 3.15

$$L_{binary} = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})] \quad (3.15)$$

Keterangan:

y : label asli (1 = *Hate Speech*, 0 = *Non-Hate Speech*)

$\hat{y}$: output prediksi model

L_{binary} : *loss binary cross-entropy*

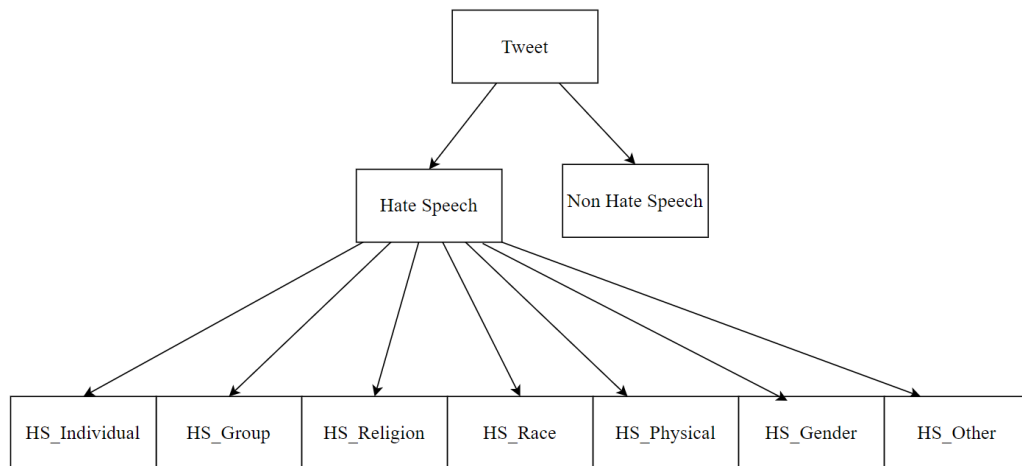
Untuk menentukan hasil klasifikasi akhir digunakan *threshold* sebesar 0.5 dimana Jika $\hat{y} > 0.5$ maka teks diklasifikasikan sebagai ujaran kebencian, sedangkan Jika $\hat{y} < 0.5$ maka teks diklasifikasikan sebagai non ujaran kebencian Perhitungan *Loss* menggunakan rumus pada 3.15 dengan label sebenarnya (*ground truth*) adalah $y = 1$ (ujaran kebencian), maka:

Model pada tahap ini dilatih menggunakan data latih dengan rasio pembagian 70:15:15. Proses pelatihan dilakukan menggunakan algoritma Adam Optimizer dengan *learning rate* sebesar $5e-5$ dan *batch size* 16. Setelah pelatihan selesai, model dievaluasi menggunakan data uji untuk menilai performa klasifikasi biner antara *Hate Speech* dan *Non-Hate Speech* berdasarkan metrik akurasi dan *F1-Score*.

3.7.2 Tahap 2: Kategori Multi-label

Pada tahap kedua, data yang telah teridentifikasi sebagai *Hate Speech* akan diklasifikasikan ke dalam satu atau lebih kategori berdasarkan target ujaran

kebencian yang muncul dalam teks. Proses ini bersifat *multi-label classification*, karena satu teks dapat mengandung lebih dari satu bentuk ujaran kebencian sekaligus. Berikut kategori dalam multi-label bisa dilihat pada Gambar 3.12



Gambar 3.12 Alur Klasifikasi Multi-Label

Representasi fitur yang digunakan sebagai input pada tahap ini berasal dari vektor [CLS] hasil *fine-tuning* model IndoBERTweet. Vektor tersebut kemudian diproses melalui *fully connected layer* dengan jumlah *neuron* sesuai banyaknya kelas (dalam penelitian ini sebanyak 7 *neuron output*). Setiap neuron menggunakan fungsi aktivasi sigmoid, yang menghasilkan probabilitas independen untuk setiap label antara 0 hingga 1. Nilai probabilitas tersebut menunjukkan seberapa besar kemungkinan suatu teks termasuk dalam masing-masing kelas ujaran kebencian. Proses prediksi multi-label dapat dijelaskan pada rumus 3.14 yang mengikuti mekanisme fungsi sigmoid sebagaimana digunakan pada tahap pertama. Nilai probabilitas dari setiap neuron kemudian dibandingkan dengan nilai *threshold* tertentu (0,5) untuk menentukan apakah label tersebut aktif atau tidak.

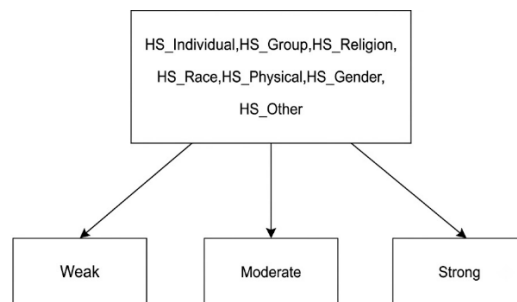
Selanjutnya, fungsi kerugian yang digunakan adalah *Binary Cross-Entropy Loss* (BCE), yang dihitung untuk setiap label dan dirata-ratakan terhadap seluruh sampel dan kelas, sebagaimana ditunjukkan pada Persamaan 3.17:

$$L_{multilabel} = -\frac{1}{N \times C} \sum_{n=1}^N \sum_{i=1}^C [y_{n,i}^{\text{true}} \log(\hat{y}_{n,i}) + (1 - y_{n,i}^{\text{true}}) \log(1 - \hat{y}_{n,i})] \quad (3.17)$$

Proses *training* dilakukan dengan optimisasi AdamW dan *Binary Cross-Entropy Loss* sebagai fungsi kerugian utama. Selama pelatihan, digunakan *early stopping* berdasarkan performa validasi untuk menghindari *overfitting*.

3.7.3 Tahap 3: Tingkat Keparahan (*Severity level*)

Tahap ketiga berfokus pada penilaian tingkat keparahan dari teks yang sudah teridentifikasi sebagai *hate speech*/ujaran kebencian pada tahap sebelumnya. Pada tahap ini, ujaran kebencian diklasifikasikan menjadi 3 tingkat yaitu *weak*, *moderate*, *strong* sebagaimana ditunjukkan pada Gambar 3.13.

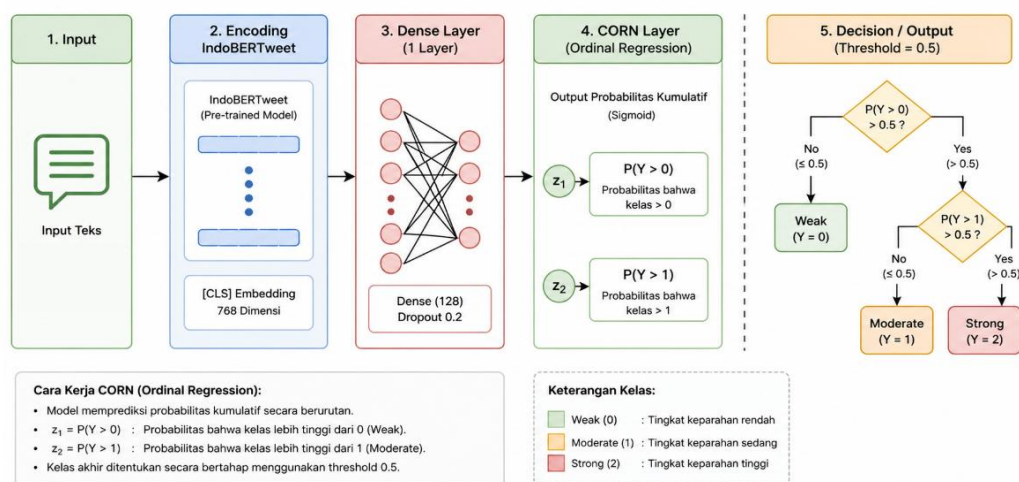


Gambar 3.13 Kategori Tingkat Keparahan

Pada Gambar 3.13 terdapat 3 tingkatan yang diperlakukan secara ordinal, karena setiap level memiliki hubungan berurutan. Model mempelajari perbedaan antar level tersebut dari representasi teks yang sama, namun dengan bobot dan pola konteks yang berbeda. Tujuan utama tahap ini adalah memberikan pemahaman

yang lebih mendalam terhadap karakteristik ujaran kebencian, tidak hanya dari sisi kategorinya, tetapi juga dari seberapa parah dampak ujaran tersebut.

Penelitian ini menggunakan pendekatan *ordinal regression* berbasis CORN (*Conditional Ordinal Regression for Neural Networks*). Pendekatan ini digunakan untuk memprediksi level keparahan yang bersifat bertingkat. CORN memecah proses prediksi menjadi keputusan biner bertahap, sehingga hasil prediksi lebih konsisten dan tetap mempertahankan urutan antar level keparahan (Shi *et al.*, 2023). Berikut Gambar 3.14 yang menampilkan arsitektur model IndoBERTweet-CORN



Gambar 3.14 Arsitektur Model dan Mekanisme Klasifikasi Menggunakan CORN

Pada Gambar 3.14 setelah melalui tahap *preprocessing*, teks diproses menggunakan model IndoBERTweet untuk memperoleh representasi fitur dari token [CLS] berukuran 768 dimensi. Vektor *embedding* tersebut kemudian diteruskan ke *dense layer* sebelum masuk ke CORN *head*.

CORN head menghasilkan dua output ordinal karena jumlah kelas sebanyak tiga $K = 3$, sehingga terbentuk $K-1 = 2$ submasalah biner, yaitu yaitu $p_1 = P(Y > 0)$ dan $p_2 = P(Y > 1)$. Output pertama merepresentasikan probabilitas bahwa tingkat

keparahan lebih tinggi dari *weak*, sedangkan output kedua merepresentasikan probabilitas bahwa tingkat keparahan lebih tinggi dari *moderate*.

Setiap output menggunakan fungsi aktivasi sigmoid untuk menghasilkan probabilitas antara 0 hingga 1 sebagaimana ditunjukkan pada persamaan 3.18

$$p_k = \sigma(z_k) = \frac{1}{1 + e^{-z_k}} \quad (3.18)$$

dengan p_k merupakan probabilitas prediksi pada *threshold* ke- k , dan z_k adalah output linier dari model.

Selanjutnya, proses pelatihan model menggunakan *Binary Cross-Entropy Loss* (BCE) pada setiap *threshold* ordinal, kemudian dijumlahkan sebagaimana pada Persamaan (3.19):

$$\mathcal{L}_{\text{CORN}} = \frac{1}{K-1} \sum_{k=1}^{K-1} \text{BCE}(p_k, b_k) \quad (3.19)$$

Dengan jumlah kelas $K=3$, maka terdapat dua submasalah biner dalam proses pelatihan model. Prediksi akhir ditentukan secara bertahap berdasarkan nilai probabilitas dengan *threshold* sebesar 0,5. Jika probabilitas pada output pertama p_1 kurang dari 0,5, maka data diklasifikasikan sebagai *weak*. Jika probabilitas pada output pertama (p_1) lebih besar atau sama dengan 0,5, namun probabilitas pada output kedua (p_2) kurang dari 0,5, maka data diklasifikasikan sebagai *moderate*. Apabila kedua probabilitas, yaitu p_1 dan p_2 , lebih besar atau sama dengan 0,5, maka data diklasifikasikan sebagai *strong*.

Dengan pendekatan ini, model tidak hanya mampu membedakan kategori ujaran kebencian, tetapi juga memahami urutan keparahan antar level, sehingga prediksi lebih konsisten dan mencerminkan realitas dampak ujaran kebencian.

3.8 Evaluasi Model

Setelah model dilatih langkah selanjutnya yaitu melakukan evaluasi seberapa baik kinerja model yang digunakan setelah melakukan *training*. Evaluasi ini bertujuan menilai kemampuan model dalam mengenali pola data dengan benar serta memastikan model dapat melakukan klasifikasi dan analisis tingkat keparahan ujaran kebencian secara konsisten. Dalam penelitian ini, proses evaluasi dibagi menjadi 3 bagian utama:

3.8.1 Evaluasi Klasifikasi

Evaluasi dilakukan untuk mengetahui sejauh mana model mampu mengklasifikasikan data ke dalam kategori ujaran kebencian (HS) dan bukan ujaran kebencian (Non-HS) secara akurat. Tahap ini bertujuan untuk mengukur performa model berdasarkan hasil prediksi terhadap data uji. Disini penelitian menggunakan *confusion matrix* untuk menganalisis distribusi model terhadap data sebenarnya. Untuk mengevaluasi model membutuhkan perhitungan metrik pada 3.20 sampai 3.23.

Akurasi merupakan metrik yang digunakan untuk menilai seberapa sering model menghasilkan prediksi yang benar, baik pada kelas positif maupun negatif.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (3.20)$$

Presisi adalah mengukur seberapa akurat prediksi positif model, yaitu seberapa banyak data yang diprediksi positif benar-benar termasuk kelas positif.

$$Precision = \frac{TP}{TP + FP} \times 100 \quad (3.21)$$

Recall merupakan metrik yang menilai sejauh mana model mampu mengenali seluruh data yang benar-benar positif, yaitu berapa banyak dari total data positif yang berhasil diprediksi dengan benar oleh model.

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (3.22)$$

F1-Score merupakan rata-rata harmoni antara nilai presisi dan *recall*, yang digunakan untuk menilai keseimbangan antara keduanya. Metrik ini sangat berguna ketika dataset bersifat tidak seimbang, sehingga mampu memberikan gambaran performa model secara lebih adil.

$$F1-Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3.23)$$

3.8.2 Evaluasi Klasifikasi Multi-label

Tahap ini bertujuan untuk mengevaluasi kinerja model dalam melakukan klasifikasi multi-label, dimana satu tweet mengandung beberapa kategori label secara bersamaan. Model yang digunakan pada tahap ini yaitu IndoBERTweet dengan fungsi aktivasi sigmoid, yang memungkinkan setiap label probabilitas independen. Oleh karena itu digunakan metrik khusus seperti *Hamming Loss*, *Macro F1-Score* dengan perhitungan sebagai berikut:

1. *Hamming Loss*

Yaitu metrik yang mengukur proporsi label yang salah diklasifikasikan terhadap total jumlah label. Semakin kecil nilai *Hamming Loss*, semakin baik performa model yang dihitung menggunakan persamaan 3.24.

$$Hamming\ Loss = \frac{1}{N \times L} \sum_{i=1}^N \sum_{j=1}^L I(y_{ij} \neq \widehat{y}_{ij}) \quad (3.24)$$

Keterangan:

N : jumlah sampel

L : jumlah label

y_{ij} : label sebenarnya (0/1)

$\widehat{y}_{ij}$: label hasil prediksi

2. *Macro F1-Score*

Macro F1-Score digunakan untuk mengukur performa model secara keseluruhan pada setiap kelas tanpa memperhatikan jumlah data di masing-masing kelas. Nilai ini diperoleh dengan menghitung rata-rata dari *F1-Score* tiap kelas, sehingga setiap kelas memiliki bobot yang sama. Berikut rumusnya ditunjukkan pada 3.25.

$$F1_{macro} = \frac{1}{L} \sum_{j=1}^L F1_j \quad (3.25)$$

3.8.3 Evaluasi Tingkat Keparahan

Evaluasi pada tahap tingkat keparahan dilakukan untuk mengukur sejauh mana model mampu menilai tingkat keparahan ujaran kebencian secara konsisten dan akurat. Karena label yang digunakan bersifat ordinal (*weak*, *moderate*, dan *strong*), maka diperlukan metrik evaluasi yang dapat mempertimbangkan urutan

atau jarak antar kategori menggunakan *Quadratic Weighted Kappa* (QWK) dan Evaluasi perlevel.

Quadratic Weighted Kappa (QWK) merupakan pengembangan dari *Cohen's Kappa* yang memberikan bobot penalti lebih besar pada kesalahan prediksi yang jauh dari kelas sebenarnya. Dengan demikian, metrik ini lebih sensitif dalam menangani perbedaan antar level keparahan. Menggunakan persamaan pada 3.28

$$\kappa_w = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}} \quad (3.28)$$

Selanjutnya yaitu matrix bobot (*Quadratic Weights*), dengan formula bobot antar kelas dengan persamaan 3.29

$$w_{i,j} = \frac{(i - j)^2}{(k - 1)^2} \quad (3.29)$$

Keterangan

$O_{i,j}$ = matriks observasi antara label aktual dan prediksi

$E_{i,j}$ = matriks ekspektasi kesepakatan secara acak

k = jumlah kelas *severity*

Pada Evaluasi per-level dilakukan untuk melihat performa model dalam mengenali setiap tingkat keparahan (*Weak, Moderate, dan Strong*). Pada tahap ini, metrik yang digunakan meliputi *Precision, Recall, dan F1-Score* untuk masing-masing level menggunakan persamaan 3.20 sampai 3.23.

3.8.4 Evaluasi E2E (*End to End*)

Evaluasi *End-to-End* (E2E) digunakan untuk menilai kinerja sistem secara menyeluruh dengan mempertimbangkan keterkaitan antar-tahap dalam *pipeline* klasifikasi. Berbeda dengan evaluasi parsial yang hanya menilai performa pada modul tunggal, pendekatan E2E menghitung performa gabungan dari seluruh

komponen sistem, mulai dari deteksi ujaran kebencian (*Hate Speech Detection*), klasifikasi kategori, hingga analisis tingkat keparahan (*Severity Level*).

Penelitian dari (Singh *et al.*, n.d.) dalam *HCDIR: End-to-End Hate Context Detection* mengintegrasikan seluruh komponen NLP dalam satu pipeline evaluasi menggunakan metrik klasifikasi gabungan, seperti *F1-Score* dan *global accuracy*. sebuah prediksi dinyatakan benar secara *end-to-end* apabila seluruh tahapan menghasilkan prediksi yang sesuai dengan label sebenarnya. Apabila terjadi kesalahan pada salah satu tahap, maka prediksi tersebut dianggap salah secara keseluruhan dalam evaluasi E2E. Evaluasi E2E hanya dilakukan pada data yang termasuk kelas ujaran kebencian ($HS = 1$), karena tahap kategori dan tingkat keparahan hanya relevan pada data tersebut. Metrik yang digunakan adalah E2E *Accuracy*, yang dirumuskan pada 3.30 sebagai berikut:

$$E2E = \frac{1}{N} \sum_{i=1}^N I(H_i = \widehat{H}_i \wedge C_i = \widehat{C}_i \wedge S_i = \widehat{S}_i) \quad (3.30)$$

Atau dapat dituliskan secara sederhana sebagai:

$$E2E = \frac{\text{jumlah data yang benar di seluruh tahap}}{\text{total data uji (HS = 1)}}$$

Keterangan:

N : jumlah total data uji

H_i, C_i, S_i : label sebenarnya

$\widehat{H}_i, \widehat{C}_i, \widehat{S}_i$: hasil prediksi model

I : fungsi indikator, bernilai 1 jika semua prediksi benar, 0 jika tidak

3.9 Skenario Uji Coba

Dalam penelitian ini, dilakukan beberapa skenario uji coba untuk mengevaluasi performa model dalam klasifikasi ujaran kebencian berbasis IndoBERTweet. Skenario yang dirancang bertujuan untuk membandingkan pengaruh pendekatan yang digunakan, yaitu pendekatan hierarkis dan *flat*, serta penerapan teknik penyeimbangan data melalui augmentasi teks. Seluruh skenario menggunakan pembagian data yang sama agar perbandingan hasil tetap adil dan terkontrol. Ringkasan konfigurasi skenario yang digunakan dalam penelitian ini disajikan pada Tabel 3.5.

Tabel 3.5 Tabel Skenario Uji Coba

Skenario	<i>Split Data</i>	<i>Balancing Data</i>	Pendekatan Klasifikasi	<i>Hyperparameter</i>
Skenario 1	70:15:15	-	Hierarkis	<i>Tokenizer: IndoBERT Tokenizer</i> <i>Maximum Sequence Length: 512</i> <i>Batch Size: 16</i> <i>Learning Rate: 5e-5</i> <i>Optimizer: AdamW</i> <i>Weight Decay (λ): 0.01</i> <i>Epoch: 5</i> <i>Scheduler: Linear Learning Rate Scheduler</i> <i>Random Seed: 42</i>
Skenario 2	70:15:15	Teks Augmentasi (MLM)	Hierarkis	
Skenario 3	70:15:15	-	Flat	

Berdasarkan Tabel 3.5, dilakukan beberapa skenario uji coba untuk mengevaluasi performa sistem klasifikasi ujaran kebencian berbasis IndoBERTweet. Setiap skenario dirancang untuk menguji perbedaan pendekatan, yaitu pendekatan hierarkis dan flat, serta pengaruh teknik penyeimbangan data melalui augmentasi teks. Seluruh skenario menggunakan pembagian data yang sama, yaitu 70:15:15. Seluruh skenario menggunakan konfigurasi *hyperparameter* yang sama untuk memastikan bahwa perbedaan performa yang dihasilkan hanya dipengaruhi oleh variasi pendekatan dan teknik penyeimbangan data.

BAB IV

HASIL DAN PEMBAHASAN

Pada bagian ini akan dijelaskan hasil dari pengujian sistem berdasarkan beberapa skenario uji coba yang telah dilakukan, dengan menggunakan beberapa evaluasi yang berbeda pada setiap tahap. Dalam mengevaluasi kinerja model yang telah dibuat, peneliti melakukan beberapa skenario pengujian dengan menggunakan penanganan kelas yang tidak seimbang, dengan menggunakan *oversampling* dengan pendekatan *text augmentation*. Seluruh pembahasan dalam bab ini mengacu pada skenario yang sudah di jelaskan pada subbab sebelumnya.

4.1 Hasil Uji Coba

Hasil uji coba dilakukan dengan 3 kali percobaan untuk mengetahui bagaimana performa dari pendekatan klasifikasi hierarkis dan flat klasifikasi. Dengan menggunakan parameter yang sama untuk setiap skenario, dan menambahkan *balancing data* pada skenario 2 untuk mengetahui peningkatan performa dalam mengatasi permasalahan ketidakseimbangan data.

4.1.1 Hasil Uji Skenario 1

Skenario pertama menggunakan parameter dengan rasio pembagian data sebesar 70:15:15, yaitu 70% *training set*, 15% *data validation set*, dan 15% *testing set*. Dataset berjumlah 11.421 data tweet yang telah melalui tahap *preprocessing* untuk menghilangkan *noise* pada teks yang tidak relevan sehingga diperoleh data bersih yang siap digunakan pada proses eksperimen.

Untuk memastikan distribusi kelas pada setiap subset data tetap seimbang, penelitian ini menggunakan teknik *stratified data splitting* dengan bantuan fungsi *train_test_split* dari *library scikit-learn*. Label pada kolom HS terlebih dahulu dikonversi menjadi bentuk biner, yaitu 0 untuk NON-HS dan 1 untuk HS, yang kemudian digunakan sebagai dasar stratifikasi. Proses pembagian data dilakukan dalam dua tahap, yaitu pembagian awal menjadi 70% data *training* dan 30% data sementara menggunakan parameter *stratify=df["label"]* yang digunakan untuk menjaga distribusi kelas pada seluruh dataset, sedangkan *stratify=y_temp* digunakan pada pembagian lanjutan untuk mempertahankan distribusi kelas pada subset data sementara agar tetap konsisten, kemudian data sementara tersebut dibagi kembali menjadi 50% *validation* dan 50% *testing* dengan tetap menerapkan stratifikasi berdasarkan *y_temp*. Dengan demikian, distribusi kelas HS dan NON-HS pada setiap subset data tetap terjaga sesuai dengan distribusi data asli, sehingga dapat mengurangi potensi bias terhadap kelas mayoritas.

Pada tahap kedua, data yang digunakan merupakan hasil filtrasi dari Tahap 1, yaitu hanya data yang diprediksi sebagai ujaran kebencian (HS). Pembagian data *training*, *validation*, dan *testing* tidak dilakukan ulang, melainkan menggunakan kembali pembagian ID dari Tahap 1 untuk menjaga konsistensi data dalam skema klasifikasi hierarkis (Guo *et al.*, 2018). Pada tahap ini dilakukan klasifikasi multi-label terhadap tujuh kategori ujaran kebencian, yaitu *HS_Religion*, *HS_Race*, *HS_Individual*, *HS_Group*, *HS_Physical*, *HS_Gender*, dan *HS_Other*, dengan menggunakan model IndoBERTweet yang dikonfigurasi untuk klasifikasi multi-label.

Pada tahap ketiga, data yang digunakan merupakan hasil lanjutan dari Tahap 2 sehingga tidak dilakukan pembagian data ulang, melainkan menggunakan kembali pembagian data sebelumnya melalui kolom split untuk menjaga konsistensi dalam skema klasifikasi hierarkis. Pada tahap ini dilakukan klasifikasi tingkat severity (keparahan) dengan memetakan label *HS_Weak*, *HS_Moderate*, dan *HS_Strong* menjadi tiga kelas ordinal, menggunakan model IndoBERTweet dengan pendekatan CORN.

Proses pelatihan digunakan model "*indolem/IndoBERTweet-base-uncased*" dengan *AutoTokenizer* IndoBERTweet, *maximum sequence length* sebesar 512, *batch size* 16, serta *learning rate* 5e-5. Proses optimasi dilakukan menggunakan AdamW optimizer dengan *weight decay* sebesar 0.01 selama 5 *epoch* pelatihan. Selain itu digunakan *linear rate scheduler* untuk mengatur penurunan *learning rate* secara bertahap selama proses *training* dan *random seed* sebesar 42 untuk memastikan reproduksibilitas eksperimen.

Penelitian ini menerapkan pendekatan klasifikasi hierarkis, sehingga jumlah data yang digunakan pada setiap tahap mengalami perubahan. Seluruh data terlebih dahulu digunakan pada Tahap 1 untuk klasifikasi HS dan Non-HS. Selanjutnya, hanya data yang diprediksi sebagai ujaran kebencian (HS) yang diteruskan ke Tahap 2 dan Tahap 3. Rincian perubahan jumlah data pada setiap tahap eksperimen ditunjukkan pada Tabel 4.1

Tabel 4.1 Rekapitulasi Dataset Pada Setiap Tahap Skenario 1

NO	Tahap Data	Jumlah Data	Training	Validation	Testing	Keterangan
1	Dataset awal	13.169	-	-	-	Data keseluruhan sebelum dilakukan <i>preprocessing</i>

NO	Tahap Data	Jumlah Data	Training	Validation	Testing	Keterangan
2	Setelah <i>preprocessing</i>	11.421	-	-	-	Setelah <i>preprocessing</i> , jumlah data menjadi 11.421 baris. Kolom <i>Abusive</i> tidak digunakan sebagai label target pada penelitian ini.
3	Dataset tahap 1	11.421	7.994	1.713	1.714	Dengan jumlah HS: 5.561 NON-HS :5.860
4	Dataset tahap 2	5.556	3.885	844	827	Data yang diprediksi HS saja yang digunakan.
5	Dataset tahap 3	5.384	3.884	749	751	Dengan jumlah data setiap level 0:3.269 1:1.645 2:470

Tabel 4.1 menunjukkan perubahan jumlah data yang digunakan pada setiap tahap dalam skenario 1. Dataset awal penelitian berjumlah 13.169 data tweet. Setelah melalui tahap *preprocessing*, jumlah data berkurang menjadi 11.421 data akibat penghapusan data yang tidak relevan. Pada penelitian ini, kolom *Abusive* juga tidak dilibatkan dalam proses klasifikasi sehingga tidak digunakan pada penelitian ini. Proses *preprocessing* tersebut membantu meningkatkan kualitas data dengan mengurangi *noise* yang tidak diperlukan.

Pada Tahap 1, seluruh data hasil *preprocessing* digunakan untuk klasifikasi awal antara ujaran kebencian (HS) dan *Non-HS*. Berdasarkan hasil klasifikasi, diperoleh 5.561 data yang diprediksi sebagai HS dan 5.860 data sebagai *Non-HS*. Distribusi ini menunjukkan bahwa data relatif seimbang, sehingga model tidak terlalu bias terhadap salah satu kelas.

Sesuai dengan pendekatan klasifikasi hierarkis, hanya data yang diprediksi sebagai HS yang diteruskan ke Tahap 2. Jumlah data yang digunakan pada tahap ini adalah 5.556 data. Terjadi sedikit pengurangan dari hasil prediksi sebelumnya,

yang mengindikasikan adanya proses *filtering* atau penyesuaian data sebelum masuk ke tahap berikutnya.

Pada Tahap 3, jumlah data kembali berkurang menjadi 5.384 data. Penurunan ini menunjukkan bahwa tidak semua data dari tahap sebelumnya dapat diproses pada tahap klasifikasi tingkat keparahan, yang kemungkinan dipengaruhi oleh proses validasi internal atau penyesuaian format data untuk model regresi ordinal

Selanjutnya evaluasi *End-to-End* (E2E) dilakukan terhadap seluruh data yang berhasil melewati ketiga tahapan klasifikasi hierarkis. Oleh karena itu, jumlah data yang digunakan pada evaluasi E2E hanya mencakup data yang diprediksi sebagai HS pada Tahap 1 dan berhasil diproses hingga Tahap 3. Hal ini menunjukkan bahwa evaluasi dilakukan secara menyeluruh terhadap *pipeline* klasifikasi dari tahap awal hingga tahap akhir. Kondisi ini mengindikasikan adanya akumulasi error atau penyaringan data pada setiap tahap, yang berdampak pada jumlah data akhir yang dievaluasi.

Pada penelitian ini, eksperimen utama dilakukan menggunakan *random seed* 42 sebagai konfigurasi utama. Untuk memastikan bahwa hasil yang diperoleh tidak bergantung pada inisialisasi acak, dilakukan percobaan tambahan sebanyak 5 kali dengan menggunakan nilai *random seed* yang berbeda dengan tujuan untuk menguji kestabilan dan konsistensi performa model, serta mengamati variasi hasil yang diperoleh sebelumnya untuk dilakukan analisis signifikan lebih lanjut. Berikut hasil rekapitulasi percobaan tersebut disajikan pada Tabel 4.2.

Tabel 4.2 Rekapitulasi Hasil Percobaan Tahap 1 Pada Skenario 1

Percobaan ke-	Precision %	Recall %	F1-Score %	Accuracy %
1	90,81	89,94	90,37	90,67
2	90,79	90,76	90,77	90,78
3	91,02	88,62	89,81	90,20
4	90,59	89,94	90,26	90,54
5	90,24	90,78	90,51	90,72
Rata-Rata (Mean)	90,69	90,01	90,34	90,58
Standar Deviasi (Std)	0,29	0,88	0,35	0,23

Berdasarkan Tabel 4.2, hasil percobaan menggunakan 5 nilai *random seed* yang berbeda menunjukkan bahwa performa model relatif konsisten pada setiap percobaan. Hal ini dapat dilihat dari nilai rata-rata (*mean*) *F1-Score* sebesar 90,34% dengan standar deviasi sebesar 0,35%, yang mengindikasikan bahwa variasi performa antar percobaan tergolong rendah

Selain itu, metrik evaluasi lainnya seperti *precision*, *recall*, dan *accuracy* juga menunjukkan nilai standar deviasi yang kecil, masing-masing sebesar 0,29%, 0,88% dan 0,23%. Nilai ini menandakan bahwa model tidak terlalu sensitif terhadap perubahan inisialisasi acak. Berdasarkan hasil tersebut, dapat disimpulkan bahwa model memiliki performa yang stabil, sehingga konfigurasi dengan *random seed* 42 dipilih sebagai representasi untuk pelaporan hasil utama pada tahap berikutnya

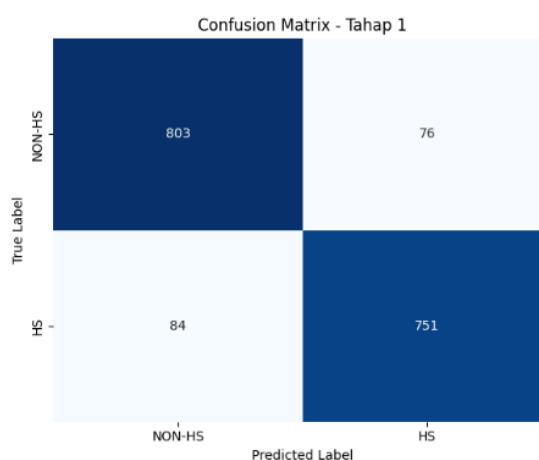
Pada Tahap 1, model menggunakan *hyperparameter max length* sebesar 512 dengan *threshold* 0,5 dan *backbone indolem/IndoBERTweet-base-uncased*. Proses pelatihan dilakukan selama 5 *epoch* dengan batch size 16 dan learning rate 5e-5. Evaluasi kinerja model dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-Score*, yang hasilnya disajikan pada Tabel 4.3.

Tabel 4.3 Hasil Evaluasi Klasifikasi Biner Tahap 1 Pada Skenario 1

	<i>Precision</i> %	<i>Recall</i> %	<i>F1-Score</i> %
0	90,53	91,35	90,94
1	90,81	89,94	90,37
<i>Accuracy</i> %	90,67		

Berdasarkan Tabel 4.3 model pada tahap 1 mampu membedakan kelas NON-HS (0) dan HS (1) dengan baik. Pada kelas NON-HS, model memperoleh *precision* 90,53%, *recall* 91,35%, dan *F1-Score* 90,94%, yang menunjukkan sebagian besar data NON-HS berhasil dikenali dengan baik.

Pada kelas HS, diperoleh *precision* 90,81%, *recall* 89,94%, dan *F1-Score* 90,37%. Hal ini menunjukkan bahwa prediksi HS cukup akurat, meskipun masih terdapat beberapa data HS yang tidak terdeteksi. Secara keseluruhan, model mencapai *accuracy* sebesar 90,67% dengan nilai *F1-Score* yang seimbang pada kedua kelas. Hal ini menunjukkan bahwa model memiliki performa yang baik dalam membedakan ujaran kebencian dan non-ujaran kebencian. Hasil prediksi tersebut selanjutnya divisualisasikan dalam bentuk *confusion matrix* pada Gambar 4.1 untuk melihat distribusi kesalahan klasifikasi yang terjadi.

Gambar 4.1 *Confusion Matrix* Klasifikasi Biner HS dan NON-HS

Confusion matrix pada Gambar 4.1, model sudah cukup baik dalam melakukan klasifikasi. Sebanyak 803 data NON-HS dan 751 data HS berhasil diprediksi dengan benar. Namun, masih ada kesalahan, yaitu 76 data NON-HS diprediksi sebagai HS dan 84 data HS diprediksi sebagai NON-HS. Secara keseluruhan, hasil prediksi yang benar lebih banyak dibandingkan yang salah, sehingga performa model dapat dikatakan baik.

Namun, error yang terjadi pada Tahap 1 ini akan terbawa ke tahap selanjutnya dalam proses klasifikasi hierarkis. Kesalahan berupa data HS yang diprediksi sebagai NON-HS menyebabkan sebagian data ujaran kebencian tidak diteruskan ke Tahap 2 dan Tahap 3, sehingga berpotensi menurunkan *recall* pada sistem secara keseluruhan. Sebaliknya, kesalahan berupa data NON-HS yang diprediksi sebagai HS akan ikut diproses pada tahap berikutnya, yang dapat menambah *noise* dan memengaruhi akurasi klasifikasi multi-label maupun penentuan tingkat keparahan.

Selanjutnya pada tahap 2 dilakukan klasifikasi multi-label untuk mengidentifikasi kategori ujaran kebencian yang terdapat dalam data. Jumlah data yang digunakan pada tahap ini sebanyak 5.556 data, yaitu data yang sebelumnya telah terklasifikasi sebagai HS pada tahap pertama. Pada tahap ini menggunakan beberapa *hyperparameter*, yaitu *max length* 512, *threshold* 0.5, *epoch* 5, *batch size* 16, dan *learning rate* 5e-5, dengan *backbone* *indolem/IndoBERTweet-base-uncased*. Pada penelitian ini dilakukan 5 kali percobaan untuk tahap 2 yang menggunakan *random seed* yang berbeda-beda dengan *seed* 42 yang digunakan

untuk analisis utama pada penelitian ini. Berikut Tabel 4.4 yang berisi rekapitulasi dari 5 kali percobaan pada tahap 2.

Tabel 4.4 Rekapitulasi Hasil Percobaan Tahap 2 Pada Skenario 1

Percobaan ke-	Macro Precision %	Macro Recall %	Macro F1-Score %	Hamming Loss %
1	76	81	78,41	9,55
2	76	81	78,38	9,47
3	77	79	77,88	9,29
4	77	83	79,59	9,47
5	77	81	78,31	9,35
Rata-Rata (Mean)	77	81	78,51	9,43
Standar Deviasi (Std)	0,55	1,41	0,64	0,10

Berdasarkan Tabel 4.4, hasil percobaan pada Tahap 2 dengan 5 nilai *random seed* yang berbeda menunjukkan bahwa performa model cenderung stabil. Hal ini terlihat dari nilai rata-rata (*mean*) *Macro F1-Score* sebesar 78,51% dengan standar deviasi sebesar 0,64%, yang mengindikasikan bahwa variasi performa antar percobaan relatif kecil.

Dari sisi *precision* dan *recall*, model menunjukkan kecenderungan nilai *recall* 81% yang lebih tinggi dibandingkan *precision* 77%. Hal ini mengindikasikan bahwa model cenderung lebih agresif dalam mendeteksi label, sehingga mampu menangkap lebih banyak label yang relevan, namun dengan konsekuensi meningkatnya kemungkinan prediksi label yang tidak tepat. Kondisi ini umum terjadi pada klasifikasi multi-label, terutama ketika terdapat overlap antar label atau distribusi label yang tidak seimbang.

Variasi yang sedikit lebih besar pada nilai *recall* (standar deviasi 1,41) dibandingkan metrik lainnya menunjukkan bahwa kemampuan model dalam menangkap seluruh label masih cukup sensitif terhadap inisialisasi acak. Hal ini

dapat disebabkan oleh kompleksitas hubungan antar label pada data multi-label, di mana perubahan kecil pada bobot awal dapat memengaruhi prediksi label tertentu.

Dari sisi kesalahan, nilai *Hamming Loss* yang rendah 9,43% dengan standar deviasi yang sangat kecil 0,10% menunjukkan bahwa kesalahan prediksi per label relatif sedikit dan konsisten. Hal ini mengindikasikan bahwa meskipun terdapat variasi pada *recall*, secara keseluruhan kesalahan prediksi model masih dapat dikendalikan dengan baik.

Secara keseluruhan, hasil ini menunjukkan bahwa model memiliki keseimbangan yang cukup baik antara kemampuan mendeteksi label (*recall*) dan ketepatan prediksi (*precision*), serta stabil terhadap variasi random seed. Dengan demikian, penggunaan *random seed* 42 sebagai konfigurasi utama dapat dianggap representatif dalam menggambarkan performa model pada tahap klasifikasi multi-label.

Pada Analisis hasil uji coba model yang menggunakan *seed* 42 menghasilkan *Macro F1-Score* sebesar 78,41% yang menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam mengenali berbagai kategori ujaran kebencian secara relatif seimbang pada setiap label. Selain itu, nilai *Hamming Loss* sebesar 9,55% yang menunjukkan bahwa tingkat kesalahan prediksi label pada model relatif rendah dalam klasifikasi multi-label. Hasil evaluasi model pada setiap kategori ujaran kebencian ditampilkan pada Tabel 4.5.

Tabel 4.5 Metrik Evaluasi Klasifikasi Multi-label Tahap 2 Pada Skenario 1

Label	<i>Precision</i> %	<i>Recall</i> %	<i>F1-Score</i> %
HS_Religion	70	85	77
HS_Race	74	86	80
HS_Individual	80	90	85
HS_Group	73	77	75

Label	<i>Precision</i> %	<i>Recall</i> %	<i>F1-Score</i> %
HS_Physical	79	69	73
HS_Gender	73	71	72
HS_Other	85	92	88
Macro Avg (%)	76	81	78,41
Hamming Loss (%)	9,5		

Berdasarkan Tabel 4.5, performa terbaik diperoleh pada kategori HS_Other dan HS_Individual dengan nilai *F1-Score* masing-masing sebesar 88% dan 85%. Hal ini menunjukkan bahwa model mampu mengenali pola ujaran kebencian pada kedua kategori tersebut dengan cukup baik. Performa yang tinggi tidak hanya dipengaruhi oleh banyaknya data namun bisa dipengaruhi oleh pola ujaran atau susunan kalimat yang lebih konsisten, sehingga lebih mudah dipelajari oleh model.

Sebaliknya, kategori seperti HS_Gender memiliki nilai *F1-Score* lebih rendah yaitu sebesar 72% yang menunjukkan bahwa model masih mengalami kesulitan dalam mengidentifikasi ujaran kebencian pada kategori tersebut. Hal ini juga dapat disebabkan oleh variasi konteks ujaran yang lebih beragam atau kemiripan dengan kategori lainnya, sehingga model lebih sulit membedakan pola pada kategori tersebut. Selain itu jumlah data pada label tersebut juga mempengaruhi nilai pada *F1-Score* tersebut, dengan adanya perbedaan performa antar kategori ini menunjukkan bahwa karakteristik data dan distribusi data dapat mempengaruhi kemampuan model dalam mempelajari pola ujaran kebencian pada klasifikasi multi-label. Distribusi jumlah data pada setiap kategori ditampilkan pada Tabel 4.6.

Tabel 4.6 Distribusi Data Pada Klasifikasi Multi-label

Rekapitulasi jumlah Distribusi Data Pada Setiap Label							
	HS_ <i>Religion</i>	HS_ <i>Race</i>	HS_ <i>Individual</i>	HS_ <i>Group</i>	HS_ <i>Physical</i>	HS_ <i>Gender</i>	HS_ <i>Other</i>
<i>Training</i>	554	391	2484	1400	219	223	2614
<i>Validation</i>	100	80	489	260	44	43	505
<i>Testing</i>	105	77	487	264	54	34	503

Perbedaan performa antar kategori pada Tabel 4.5 selanjutnya dapat dianalisis melalui distribusi data pada setiap label yang ditampilkan pada Tabel 4.6. Berdasarkan tabel tersebut, terlihat bahwa jumlah data pada setiap label tidak seimbang. Label seperti *HS_Other* dan *HS_Individual* memiliki jumlah data yang lebih banyak dibandingkan label lainnya, seperti *HS_Gender* dan *HS_Physical*. Hal ini sejalan dengan hasil evaluasi pada Tabel 4.5 di mana kategori dengan jumlah data yang lebih besar cenderung memiliki performa yang lebih tinggi. Sebaliknya, kategori dengan jumlah data yang lebih sedikit menunjukkan performa yang lebih rendah. Hal ini mengindikasikan bahwa keterbatasan jumlah data mempengaruhi kemampuan model dalam mempelajari pola ujaran kebencian pada kategori tersebut. Berikut rekapitulasi data jumlah label dalam 1 tweet pada Tabel 4.7.

Tabel 4.7 Rekapitulasi Jumlah Label Dalam 1 Tweet Pada Skenario 1

Rekapitulasi Jumlah Label Dalam 1 Tweet							
	1	2	3	4	5	6	7
<i>Training</i>	-	3771	109	4	-	-	-
<i>Validation</i>	-	728	19	2	-	-	-
<i>Testing</i>	-	730	20	1	-	-	-

Pada Tabel 4.7 persebaran jumlah label dalam satu tweet juga tidak seimbang. Sebagian besar tweet memiliki dua label, sedangkan tweet dengan jumlah label lebih banyak, seperti empat label, jumlahnya sangat sedikit. Kondisi ini menunjukkan bahwa distribusi kombinasi label pada data cenderung tidak merata. Ketidakseimbangan distribusi data ini juga berpengaruh terhadap

penentuan *threshold* dalam proses klasifikasi. Dalam penelitian ini, penggunaan *threshold* sebesar 0,5 dinilai cukup efektif untuk membantu model dalam menentukan label prediksi secara seimbang antara *precision* dan *recall*. Selanjutnya pada Tabel 4.8 menyajikan bagaimana model melakukan prediksi serta kesalahan yang dihasilkan pada setiap kategori, dengan dilakukan analisis menggunakan *confusion matrix* multi-label per label.

Tabel 4.8 *Confusion Matrix* Multi-label Tahap 2 Pada Skenario 1

label	TP	FP	FN	TN
HS_Religion	89	38	16	684
HS_Race	66	23	11	727
HS_Individual	440	112	47	228
HS_Group	203	75	61	488
HS_Physical	37	10	17	763
HS_Gender	24	9	10	784
HS_Other	462	83	41	241

Berdasarkan Tabel 4.8, label HS_Other TP = 462 dan HS_Individual TP = 440 memiliki nilai *True Positive* tertinggi dibandingkan label lainnya. Hal ini menunjukkan bahwa model mampu mengenali kedua kategori tersebut dengan baik. Tingginya nilai TP ini kemungkinan dipengaruhi oleh jumlah data yang lebih besar serta pola ujaran yang lebih eksplisit, sehingga lebih mudah dipelajari oleh model.

Namun, pada label HS_Group terlihat nilai *False Negative* yang cukup tinggi FN = 61, yang menunjukkan bahwa masih banyak data yang seharusnya termasuk dalam kategori tersebut tetapi tidak berhasil dideteksi. Hal ini mengindikasikan bahwa ujaran berbasis kelompok cenderung memiliki konteks yang lebih kompleks atau ambigu, sehingga sulit dibedakan dari kategori lain dalam skenario multi-label.

Selain itu, nilai *False Positive* yang cukup tinggi juga terlihat pada label *HS_Individual* $FP = 112$ dan *HS_Religion* $FP = 38$. Hal ini menunjukkan bahwa model cenderung melakukan *over-prediction* pada label tersebut, yaitu mengklasifikasikan data ke dalam kategori tertentu meskipun sebenarnya tidak termasuk. Kesalahan ini kemungkinan disebabkan oleh adanya kemiripan konteks atau tumpang tindih makna antar label.

Di sisi lain, label seperti *HS_Physical* $TP = 37$ dan *HS_Gender* $TP = 24$ memiliki nilai *true positive* yang relatif rendah. Hal ini menunjukkan bahwa kemampuan model dalam mendeteksi data positif pada kedua kategori tersebut masih terbatas. Kondisi ini kemungkinan dipengaruhi oleh ketidakseimbangan jumlah data, di mana kelas dengan jumlah data lebih sedikit cenderung lebih sulit dipelajari oleh model.

Secara keseluruhan, hasil *confusion matrix* ini menunjukkan bahwa performa model dalam klasifikasi multi-label dipengaruhi oleh jumlah data, kompleksitas konteks bahasa, serta adanya tumpang tindih antar kategori. Selain itu, ketidakseimbangan distribusi data juga berdampak pada rendahnya kemampuan model dalam mendeteksi kelas minoritas. Oleh karena itu, diperlukan strategi tambahan seperti penyeimbangan data atau penyesuaian *threshold* untuk meningkatkan performa model.

Selanjutnya pada tahap 3 dilakukan pengujian model untuk memprediksi tingkat keparahan (*severity*) ujaran kebencian menjadi tiga level, yaitu *weak*, *moderate*, dan *strong*. Pengujian ini dilakukan pada data yang telah diproses pada

tahap sebelumnya dengan total 5.384 data, yang terdiri dari 3.269 data *weak*, 1.645 data *moderate*, dan 470 data *strong*.

Evaluasi model pada tahap ini menggunakan *Quadratic Weighted Kappa* (QWK) untuk mengukur tingkat kesesuaian antara prediksi model dan label sebenarnya dengan mempertimbangkan sifat ordinal pada tingkat keparahan. Dengan demikian, kesalahan prediksi antar level tidak memiliki bobot yang sama, di mana kesalahan yang lebih jauh akan diberikan penalti yang lebih besar. Pada tahap ini menggunakan beberapa *hyperparameter*, yaitu *max length* 512, *threshold* 0.5, *epoch* 5, *batch size* 16, dan *learning rate* 5e-5, dengan *backbone* *indolem/IndoBERTweet-base-uncased*, dengan menggunakan pendekatan CORN (*Conditional Ordinal Regression for Neural networks*) untuk menangani permasalahan klasifikasi ordinal. Pada penelitian ini juga melakukan 5 kali percobaan seperti pada tahap 1 dan 2 dengan menggunakan *seed* yang berbeda setiap 1 kali percobaan. Namun untuk analisis utama pada penelitian ini menggunakan *seed* 42. Berikut rekapitulasi dari beberapa kali percobaan pada Tabel 4.9.

Tabel 4.9 Rekapitulasi Hasil Percobaan Tahap 3 Pada Skenario 1

Percobaan ke-	Macro Precision %	Macro Recall %	Macro F1-Score %	Accuracy	QWK %
1	79,06	82,28	80,40	83,22	75,57
2	79,98	80,08	80,01	83,36	72,79
3	85,75	85,48	85,57	86,35	81,23
4	81,49	79,67	80,52	82,02	74,57
5	83,31	80,05	81,55	82,71	74,84
Rata-Rata (Mean)	81,92	81,51	81,61	83,53	75,80
Standar Deviasi (Std)	2,68	2,45	2,29	1,66	3,20

Berdasarkan Tabel 4.9, hasil percobaan pada Tahap 3 menunjukkan bahwa performa model cenderung lebih bervariasi dibandingkan tahap sebelumnya. Hal ini terlihat dari nilai rata-rata (*mean*) *Macro F1-Score* sebesar 81,61% dengan standar deviasi sebesar 2,29%, yang menunjukkan adanya variasi performa yang cukup signifikan antar percobaan.

Variasi yang relatif tinggi juga terlihat pada metrik *macro precision* 2,68% dan *macro recall* 2,45%, yang mengindikasikan bahwa model cukup sensitif terhadap perubahan inisialisasi acak pada tahap ini. Hal ini dapat disebabkan oleh kompleksitas tugas klasifikasi tingkat keparahan (*ordinal*), di mana model tidak hanya menentukan label, tetapi juga harus memahami urutan atau tingkatan antar kelas.

Selain itu, perbedaan performa yang cukup mencolok pada percobaan ke 3 dibandingkan percobaan lainnya menunjukkan bahwa inisialisasi tertentu dapat menghasilkan konfigurasi model yang lebih optimal. Hal ini mengindikasikan bahwa proses pelatihan pada tahap ini memiliki ketergantungan yang lebih tinggi terhadap kondisi awal dibandingkan tahap sebelumnya.

Dari sisi evaluasi *ordinal*, nilai QWK memiliki rata-rata sebesar 75,80% dengan standar deviasi tertinggi, yaitu 3,20%. Hal ini menunjukkan bahwa konsistensi model dalam menangkap hubungan *ordinal* antar kelas masih belum sepenuhnya stabil. Variasi ini mengindikasikan bahwa pemahaman model terhadap perbedaan tingkat keparahan masih dipengaruhi oleh faktor inisialisasi dan distribusi data pada masing-masing kelas.

Meskipun demikian, nilai rata-rata *accuracy* sebesar 83,53% dengan standar deviasi 1,66% menunjukkan bahwa secara umum model tetap mampu melakukan klasifikasi dengan baik. Namun, dibandingkan dengan tahap sebelumnya, hasil ini menunjukkan bahwa tahap klasifikasi tingkat keparahan merupakan bagian yang lebih menantang dalam pipeline hierarkis.

Secara keseluruhan, hasil ini menunjukkan bahwa model pada Tahap 3 memiliki performa yang cukup baik, namun dengan tingkat kestabilan yang lebih rendah dibandingkan tahap sebelumnya. Dengan demikian, penggunaan *random seed* 42 sebagai konfigurasi utama tetap digunakan, namun perlu diperhatikan bahwa variasi performa pada tahap ini lebih tinggi dan berpotensi memengaruhi hasil akhir sistem secara keseluruhan. Hasil evaluasi pada setiap tingkat keparahan selanjutnya disajikan pada Tabel 4.10.

Tabel 4.10 Hasil Evaluasi Tingkat Keparahan Tahap 3 Pada Skenario 1

Severity	Precision %	Recall %	F1-Score %
Weak	87,99	88,57	88,28
Moderate	77,25	71,49	74,26
Strong	71,95	86,76	78,67
Accuracy	83,22		
Quadratic Weighted Kappa (QWK)	75,57		

Berdasarkan Tabel 4.10, performa model menunjukkan variasi yang cukup signifikan antar tingkat keparahan. Pada kategori *weak*, model mencapai *F1-Score* sebesar 88,28% dengan nilai *precision* dan *recall* yang relatif seimbang. Hal ini menunjukkan bahwa model mampu mengenali ujaran dengan tingkat keparahan rendah secara konsisten, baik dalam meminimalkan kesalahan prediksi positif maupun dalam menangkap sebagian besar data yang relevan. Performa yang tinggi

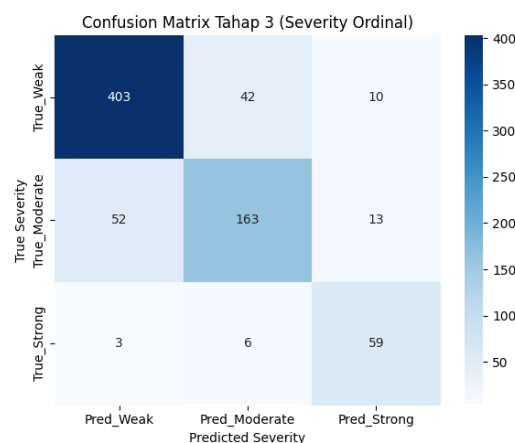
pada kategori ini kemungkinan dipengaruhi oleh distribusi data yang lebih dominan, sehingga model memiliki representasi pola yang lebih baik.

Pada kategori *strong*, model menunjukkan *recall* yang tinggi 86,76%, namun *precision* relatif lebih rendah 71,95%. Kondisi ini mengindikasikan bahwa model cenderung lebih agresif dalam mengklasifikasikan ujaran sebagai kategori *strong*, sehingga mampu menangkap sebagian besar data aktual, tetapi juga meningkatkan kemungkinan kesalahan prediksi *false positive*. Fenomena ini menunjukkan adanya kecenderungan model untuk *over-predict* pada kelas dengan tingkat keparahan tinggi.

Sebaliknya, kategori *moderate* menjadi kelas dengan performa terendah, dengan *F1-Score* sebesar 74,26%. Nilai *recall* yang lebih rendah dibandingkan *precision* menunjukkan bahwa model masih kesulitan dalam mengidentifikasi seluruh data pada kategori ini. Hal ini dapat disebabkan oleh sifat kategori *moderate* yang berada di antara *weak* dan *strong*, sehingga memiliki karakteristik yang lebih ambigu dan tumpang tindih dengan kedua kelas lainnya. Akibatnya, model cenderung mengalami kebingungan dalam menentukan batas keputusan antar kelas.

Secara keseluruhan, model memperoleh *accuracy* sebesar 83,22% dan nilai QWK sebesar 75,57%, yang menunjukkan tingkat kesesuaian yang cukup baik antara prediksi model dan label aktual dalam konteks klasifikasi ordinal. Nilai QWK yang relatif tinggi mengindikasikan bahwa meskipun masih terdapat kesalahan prediksi, sebagian besar kesalahan tersebut terjadi pada kelas yang berdekatan (misalnya *moderate* diprediksi sebagai *weak*). Hal ini menunjukkan bahwa model mampu mempertahankan hubungan ordinal antar tingkat keparahan.

Penelitian ini menunjukkan bahwa pendekatan CORN cukup efektif dalam memodelkan klasifikasi ordinal karena mampu memecah permasalahan menjadi keputusan bertahap. Dengan pendekatan ini, model menghasilkan prediksi yang lebih terstruktur dan cenderung menghindari kesalahan yang melompati lebih dari satu tingkat keparahan. Namun, masih terdapat tantangan dalam membedakan kategori dengan batas yang tidak tegas, khususnya pada kelas *moderate*. Oleh karena itu, diperlukan strategi tambahan seperti penyeimbangan data atau eksplorasi *threshold* yang lebih optimal untuk meningkatkan performa pada kelas tersebut. Untuk lebih jelasnya berikut Gambar 4.2 untuk *confusion matrix* dari tahap 3 tingkat keparahan.



Gambar 4.2 *Confusion Matrix* Tingkat Keparahannya Pada Tahap 3 Skenario 1

Berdasarkan Gambar 4.2, sebagian besar prediksi model berada pada diagonal utama, yang menunjukkan bahwa model sudah mampu mengklasifikasikan tingkat keparahan dengan cukup baik. Namun, masih terdapat beberapa kesalahan yang memberikan gambaran mengenai kelemahan model.

Pada kategori *weak*, sebagian besar data berhasil diprediksi dengan benar. Kesalahan yang terjadi umumnya bergeser ke kategori *moderate*, sehingga

menunjukkan bahwa model masih mengalami sedikit kebingungan dalam membedakan batas antara kedua kelas tersebut.

Pada kategori *moderate*, kesalahan prediksi terlihat lebih banyak dibandingkan kelas lainnya. Sebagian data diprediksi sebagai *weak*, yang menunjukkan bahwa model cenderung menurunkan tingkat keparahan pada kategori ini. Hal ini mengindikasikan bahwa kelas *moderate* masih sulit dibedakan karena posisinya berada di antara dua kelas lainnya.

Pada kategori *strong*, model mampu mengenali sebagian besar data dengan baik. Kesalahan yang terjadi relatif sedikit dan sebagian besar masih berada pada kelas yang berdekatan, yaitu *moderate*.

Temuan ini memperkuat hasil evaluasi sebelumnya, di mana kategori *moderate* menjadi kelas yang paling sulit diprediksi. Hal ini disebabkan oleh sifatnya yang berada di tengah (*ambiguous boundary*), sehingga memiliki karakteristik yang tumpang tindih dengan kedua kelas lainnya. Selain itu, distribusi kesalahan yang dominan pada kelas berdekatan menunjukkan bahwa pendekatan ordinal menggunakan CORN telah bekerja dengan baik dalam menjaga struktur urutan label. Hal ini juga didukung oleh pola kesalahan pada *confusion matrix* yang sebagian besar terjadi pada kelas yang berdekatan, sehingga menunjukkan bahwa model telah cukup baik dalam memahami hubungan ordinal antar tingkat keparahan.

Selanjutnya dilakukan evaluasi *end-to-end* (E2E) untuk menilai kinerja sistem secara keseluruhan dengan mengintegrasikan hasil prediksi dari Tahap 1, Tahap 2, dan Tahap 3 dalam satu pipeline klasifikasi. Evaluasi ini bertujuan untuk

mengukur kemampuan sistem dalam menghasilkan prediksi yang benar ketika seluruh tahapan dijalankan secara berurutan, mengingat kesalahan pada satu tahap dapat memengaruhi tahap berikutnya.

Evaluasi dilakukan menggunakan data uji ujaran kebencian (HS) yang diproses secara berurutan melalui tiga tahap, yaitu deteksi HS pada Tahap 1, klasifikasi kategori secara multi-label pada Tahap 2, serta klasifikasi tingkat keparahan pada Tahap 3.

Pada Tahap 1, prediksi dianggap benar apabila label hasil prediksi sesuai dengan label aktual. Data yang salah diklasifikasikan sebagai Non-HS tidak diteruskan ke tahap berikutnya dan dicatat sebagai kesalahan. Sementara itu, pada Tahap 2 dan Tahap 3, evaluasi menggunakan pendekatan subset, di mana prediksi dianggap benar apabila label yang diprediksi sesuai dengan label sebenarnya.

Dalam penelitian ini digunakan dua pendekatan evaluasi E2E, yaitu *E2E Evaluation System* dan *E2E Evaluation HS-only*. Pendekatan *E2E System* digunakan untuk mengevaluasi kinerja sistem secara menyeluruh sepanjang *pipeline*, sedangkan *HS-only* digunakan untuk mengevaluasi performa sistem dengan berfokus pada data ujaran kebencian, sehingga kesalahan pada setiap tahap dapat dianalisis lebih spesifik.

Selain itu, dilakukan lima kali percobaan dengan menggunakan nilai random seed yang berbeda, serupa dengan Tahap 1 hingga Tahap 3 untuk menguji kestabilan hasil evaluasi E2E. Dengan *seed* 42 menjadi *baseline* utama untuk dilakukan analisis, hasil dari setiap percobaan tersebut disajikan pada Tabel 4.11.

Tabel 4.11 Rekapitulasi Hasil Percobaan E2E Pada Skenario 1

Percobaan ke-	E2E HS %	E2E System %
1	74,83	79,63
2	72,96	78,93
3	74,72	79,28
4	75,23	79,69
5	74,93	79,63
Rata-Rata (Mean)	74,53	79,43
Standar Deviasi (Std)	0,90	0,32

Berdasarkan Tabel 4.11, hasil evaluasi *end-to-end* (E2E) menunjukkan bahwa performa sistem secara keseluruhan relatif stabil pada setiap percobaan. Hal ini terlihat dari nilai rata-rata (*mean*) E2E System sebesar 79,43% dengan standar deviasi yang rendah, yaitu 0,32, yang mengindikasikan bahwa variasi performa antar percobaan sangat kecil.

Sementara itu, nilai E2E HS memiliki rata-rata sebesar 74,53% dengan standar deviasi sebesar 0,90, yang menunjukkan variasi yang sedikit lebih tinggi dibandingkan E2E System. Hal ini mengindikasikan bahwa performa sistem dalam menangani data ujaran kebencian (HS) lebih sensitif terhadap perubahan inisialisasi acak dibandingkan performa sistem secara keseluruhan.

Perbedaan nilai antara E2E System dan E2E HS menunjukkan adanya dampak dari proses klasifikasi bertingkat (*pipeline*), di mana kesalahan pada tahap 1 dapat menyebabkan sebagian data HS tidak diteruskan ke tahap berikutnya. Hal ini berkontribusi terhadap penurunan performa pada evaluasi E2E HS. Selain itu, kesalahan pada Tahap 2 dan Tahap 3 juga dapat terakumulasi dan memengaruhi hasil akhir sistem.

Meskipun demikian, nilai standar deviasi yang relatif kecil pada kedua metrik menunjukkan bahwa sistem memiliki kestabilan yang baik terhadap variasi random seed. Dengan demikian, hasil evaluasi E2E dapat dianggap konsisten dan representatif dalam menggambarkan performa keseluruhan sistem klasifikasi hierarkis yang dikembangkan. Untuk memahami lebih lanjut bagaimana kesalahan terdistribusi pada setiap tahap dalam pipeline klasifikasi, dilakukan analisis terhadap hasil prediksi pada masing-masing tahap. Analisis ini bertujuan untuk mengidentifikasi pada tahap mana terjadi penurunan performa yang paling signifikan dalam proses klasifikasi hierarkis. Distribusi hasil prediksi pada setiap tahap menggunakan *seed* 42 ditunjukkan pada Tabel 4.12.

Tabel 4.12 Hasil Distribusi Status E2E Pada Setiap Tahap Skenario 1

<i>E2E HS Only</i>		<i>E2E System</i>	
Jumlah <i>testing set</i>	1,714	Jumlah <i>testing set</i>	1.714
Jumlah yang terdeteksi HS	751	Jumlah benar NON-HS	803
Jumlah benar tahap 1	751	Jumlah benar tahap 1	751
Jumlah benar tahap 2	570	Jumlah benar tahap 2	570
Jumlah benar tahap 3	562	Jumlah benar tahap 3	562
E2E Akurasi (%)	74,83%	E2E Akurasi (%)	79,63%

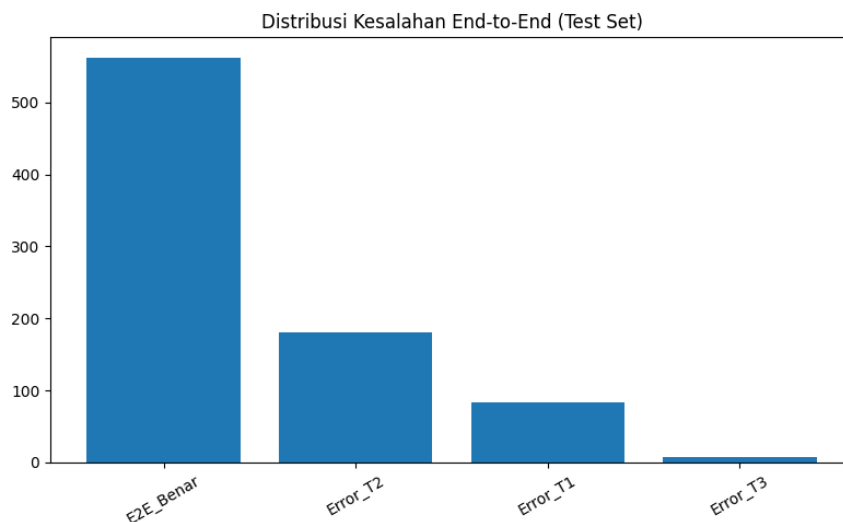
Berdasarkan Tabel 4.12, evaluasi *end-to-end* dilakukan menggunakan dua pendekatan, yaitu *E2E evaluation HS only* dan *E2E evaluation System*. Pada pendekatan *E2E System*, evaluasi mengikuti alur sistem, di mana hanya data yang diprediksi sebagai ujaran kebencian pada Tahap 1 yang diteruskan ke tahap berikutnya. Dari 751 data HS yang masuk ke *pipeline*, sebanyak 562 data berhasil diprediksi dengan benar ditambah 803 data yang di prediksi benar NON-HS, sehingga diperoleh *E2E Accuracy* sebesar 79,63%. Hasil ini menunjukkan bahwa ketika data berhasil melewati tahap deteksi awal, model pada tahap berikutnya mampu bekerja dengan cukup baik.

Sementara itu, pada pendekatan *E2E HS Only*, evaluasi dilakukan terhadap seluruh 751 data ujaran kebencian berdasarkan label aktual. Dari jumlah tersebut, sebanyak 562 data berhasil diprediksi dengan benar, dengan *E2E Accuracy* sebesar 74,83% nilai yang lebih rendah menunjukkan bahwa kesalahan pada Tahap 1 memberikan dampak signifikan terhadap performa sistem secara keseluruhan.

Distribusi kesalahan menunjukkan bahwa kesalahan terbesar terjadi pada Tahap 2, yaitu sebanyak 181 data. Hal ini menunjukkan bahwa klasifikasi multi-label masih menjadi tantangan utama, terutama karena satu data dapat memiliki lebih dari satu kategori serta adanya kemiripan antar kategori yang dapat menyebabkan ambiguitas.

Pada Tahap 1, terdapat 84 data yang salah diklasifikasikan sebagai Non-HS, sehingga tidak dapat diteruskan ke tahap berikutnya. Sementara itu, kesalahan pada Tahap 3 relatif kecil, yaitu hanya 8 data. Hal ini menunjukkan bahwa model klasifikasi tingkat keparahan cukup stabil ketika kategori ujaran kebencian telah berhasil diidentifikasi.

Secara keseluruhan, hasil ini menunjukkan bahwa performa sistem *end-to-end* sangat dipengaruhi oleh akumulasi kesalahan antar tahap, terutama pada Tahap 2 yang memiliki kompleksitas lebih tinggi. Meskipun demikian, pendekatan ini masih mampu menghasilkan performa yang lebih baik ketika tahap awal berhasil menyaring data dengan tepat. Pada Gambar 4.3 menunjukkan distribusi kesalahan pada evaluasi *end-to-end* skenario 1.



Gambar 4.3 Distribusi Kesalahan E2E Untuk Skenario 1

Visualisasi pada Gambar 4.3 menunjukkan bahwa kinerja sistem *end-to-end* dipengaruhi oleh akumulasi kesalahan antar tahap dalam *pipeline*. Meskipun setiap model memiliki performa yang cukup baik secara individual, integrasi dalam satu alur menyebabkan kesalahan pada satu tahap berdampak pada hasil akhir sistem.

Kesalahan terbesar terlihat pada Tahap 2 (klasifikasi multi-label), sehingga menjadi faktor utama penurunan kinerja secara keseluruhan. Hal ini menunjukkan bahwa peningkatan performa pada tahap tersebut berpotensi memberikan dampak paling signifikan terhadap kinerja sistem *end-to-end*.

4.1.2 Hasil Uji Skenario 2

Pada skenario 2, model klasifikasi kategori tetap namun perbedaan utama pada skenario ini terletak pada penerapan penanganan *class imbalance* melalui teknik oversampling berbasis *text augmentation* menggunakan pendekatan *Masked Language Modeling* (MLM). Teknik tersebut diterapkan pada Tahap 2 (klasifikasi multi-label) karena distribusi data antara kategori ujaran kebencian tidak seimbang.

Sementara itu, pada Tahap 1 distribusi data antara kelas HS dan Non-HS relatif seimbang sehingga tidak memerlukan penanganan khusus. Pada Tahap 3, penanganan *class imbalance* tidak diterapkan untuk menjaga konsistensi representasi data yang bersifat ordinal pada klasifikasi tingkat keparahan. Rincian perubahan jumlah data pada setiap tahap eksperimen ditunjukkan pada Tabel 4.13.

Tabel 4.13 Rekapitulasi Dataset Pada Setiap Tahap Skenario 2

NO	Tahap Data	Jumlah Data	Training	Validation	Testing	Keterangan
1	Dataset awal	13.169	-	-	-	Data keseluruhan sebelum dilakukan <i>preprocessing</i>
2	Setelah <i>preprocessing</i>	11.421	-	-	-	Setelah <i>preprocessing</i> , jumlah data menjadi 11.421 baris. Kolom <i>Abusive</i> tidak digunakan sebagai label target pada penelitian ini.
3	Dataset tahap 1	11.421	7.994	1.713	1.714	Dengan jumlah HS: 5.561 NON-HS :5.860
4	Dataset tahap 2 (sebelum MLM)	5.556	3.885	844	827	Data yang diprediksi HS saja yang digunakan.
5	Dataset tahap 2 (setelah MLM)	6.386	4.715	844	827	Data prediksi HS dengan augmentasi MLM pada <i>training</i>
5	Dataset tahap 3	5.378	3.884	743	751	Dengan jumlah data setiap level 0:3.259 1:1.648 2:471

Tabel 4.13 menunjukkan proses pembentukan dataset pada eksperimen skenario 2 yang menggunakan pendekatan klasifikasi hierarkis. Dataset awal berjumlah 13.169 data, kemudian setelah tahap *preprocessing* menjadi 11.421 data. Pada penelitian ini, kolom *Abusive* tidak digunakan sebagai label target.

Pada Tahap 1, dilakukan klasifikasi biner antara HS dan NON-HS menggunakan seluruh dataset hasil *preprocessing*. Dari proses ini diperoleh 5.561

data HS yang kemudian diteruskan ke tahap berikutnya. Hal ini menunjukkan bahwa hanya data yang terdeteksi sebagai ujaran kebencian yang akan diproses lebih lanjut pada klasifikasi kategori dan tingkat keparahan. Selanjutnya pada Tahap 2, klasifikasi dilakukan menggunakan 5.556 data HS untuk mengidentifikasi kategori ujaran kebencian secara multi-label. Pada tahap ini diterapkan teknik *oversampling* berbasis *text augmentation* untuk mengatasi ketidakseimbangan distribusi data antar kategori, sehingga jumlah data pada *training set* bertambah, sebanyak 867 data label tambahan.

Pada Tahap 3, data HS digunakan untuk memprediksi tingkat keparahan (*severity*) yang terdiri dari tiga level, yaitu *weak* (0) sebanyak 3.231 data, *moderate* (1) sebanyak 1.641 data, dan *strong* (2) sebanyak 471 data. Distribusi ini menunjukkan bahwa kelas *weak* mendominasi dataset dibandingkan level lainnya. Selanjutnya, evaluasi *End-to-End* (E2E) dilakukan terhadap data uji HS yang diproses melalui seluruh tahapan klasifikasi hierarkis, mulai dari deteksi ujaran kebencian hingga prediksi tingkat keparahan. Pada skenario 2 ini dilakukan 5 kali percobaan menggunakan *seed* yang berbeda, berikut rekapitulasi hasil dari beberapa percobaan pada Tabel 4.14.

Tabel 4.14 Rekapitulasi Hasil Percobaan Tahap 1 Pada Skenario 2

Percobaan ke-	<i>Precision</i> %	<i>Recall</i> %	<i>F1-Score</i> %	<i>Accuracy</i> %
1	90,81	89,94	90,37	90,67
2	91,03	89,94	90,48	90,78
3	91,06	89,10	90,07	90,43
4	90,59	89,94	90,26	90,55
5	90,23	90,77	90,50	90,72
Rata-Rata (Mean)	90,74	89,94	90,34	90,63
Standar Deviasi (Std)	0,34	0,59	0,18	0,14

Berdasarkan Tabel 4.14, hasil percobaan pada Tahap 1 dengan lima nilai *random seed* yang berbeda menunjukkan bahwa performa model sangat stabil. Hal ini terlihat dari nilai rata-rata (*mean*) *F1-Score* sebesar 90,34% dengan standar deviasi yang sangat kecil, yaitu 0,18%, yang mengindikasikan bahwa variasi performa antar percobaan hampir tidak signifikan. Metrik evaluasi lainnya seperti *precision*, *recall*, dan *accuracy* juga menunjukkan konsistensi yang tinggi, dengan nilai standar deviasi masing-masing sebesar 0,34%, 0,59%, dan 0,14%. Nilai ini menunjukkan bahwa model tidak sensitif terhadap perubahan inisialisasi acak pada tahap klasifikasi biner.

Nilai *precision* yang relatif tinggi dan stabil menunjukkan bahwa model mampu meminimalkan kesalahan dalam memprediksi data NON-HS sebagai HS. Sementara itu, nilai *recall* yang sedikit lebih rendah dan memiliki variasi lebih besar dibandingkan metrik lainnya menunjukkan bahwa masih terdapat beberapa data HS yang tidak terdeteksi oleh model. Hal ini sejalan dengan hasil *confusion matrix* pada Gambar 4.1 tahap 1 skenario 1, di mana terdapat sejumlah data HS yang salah diklasifikasikan sebagai NON-HS.

Meskipun demikian, perbedaan antara *precision* dan *recall* tidak terlalu besar, sehingga menghasilkan nilai *F1-Score* yang tetap tinggi dan seimbang. Hal ini menunjukkan bahwa model memiliki kemampuan yang baik dalam membedakan antara kelas HS dan NON-HS secara keseluruhan.

Secara keseluruhan, hasil ini menunjukkan bahwa model pada Tahap 1 memiliki performa yang tidak hanya tinggi, tetapi juga stabil terhadap variasi *random seed*. Dengan demikian, penggunaan *random seed* 42 sebagai konfigurasi

utama dapat dianggap representatif dalam menggambarkan performa model pada tahap klasifikasi biner.

Selanjutnya analisis pada Tahap 1, seluruh data hasil *preprocessing* digunakan untuk melakukan klasifikasi biner antara ujaran kebencian (HS) dan non-ujaran kebencian Non-HS. Model dikonfigurasi dengan *maximum sequence length* sebesar 512, serta menggunakan *backbone indolem/IndoBERTweet-base-uncased*. Proses pelatihan dilakukan selama 5 *epoch* dengan *batch size* 16 dan *learning rate* 5e-5.

Evaluasi kinerja model dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-Score* untuk mengukur performa klasifikasi pada masing-masing kelas. Hasil evaluasi tersebut disajikan pada Tabel 4.15.

Tabel 4.15 Hasil Evaluasi Klasifikasi Biner Tahap 1 Pada Skenario 2

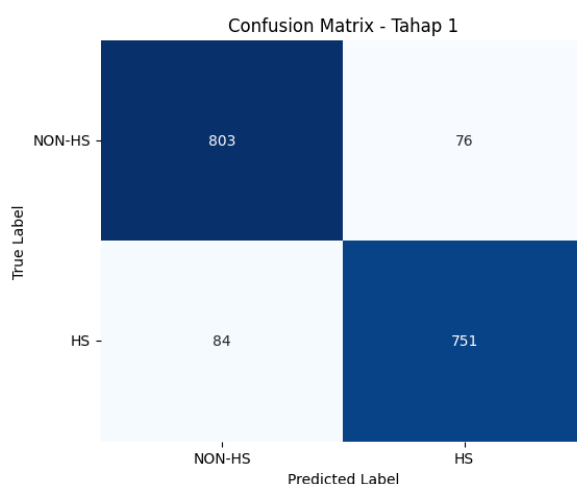
	<i>Precision</i> %	<i>Recall</i> %	<i>F1-Score</i> %
0	90,53	91,35	90,94
1	90,81	89,94	90,37
<i>Accuracy</i> (%)	90,67%		

Berdasarkan Tabel 4.15, model pada Tahap 1 menunjukkan performa yang baik dan relatif seimbang antara kelas NON-HS (0) dan HS (1). Kedua kelas memiliki nilai *precision*, *recall*, dan *F1-Score* yang hampir setara, dengan *F1-Score* masing-masing sebesar 90,94% untuk NON-HS dan 90,37% untuk HS. Hal ini menunjukkan bahwa model mampu menjaga keseimbangan antara kemampuan mendeteksi data positif dan meminimalkan kesalahan prediksi.

Model juga mencapai *accuracy* sebesar 90,67%, yang menunjukkan bahwa sebagian besar data berhasil diklasifikasikan dengan benar. Meskipun demikian,

nilai recall pada kelas HS sedikit lebih rendah dibandingkan NON-HS, yang mengindikasikan bahwa masih terdapat beberapa ujaran kebencian yang belum terdeteksi.

Secara keseluruhan, hasil ini menunjukkan bahwa model cukup andal sebagai tahap deteksi awal dalam pipeline klasifikasi, sebelum dilanjutkan ke tahap klasifikasi kategori dan tingkat keparahan. Visualisasi distribusi prediksi model ditunjukkan melalui *confusion matrix* pada Gambar 4.4.



Gambar 4.4 *Confusion Matrix* Klasifikasi Biner HS dan NON-HS

Berdasarkan Gambar 4.4, sebagian besar data berhasil diklasifikasikan dengan benar oleh model. Pada kelas NON-HS, terdapat 803 data yang diprediksi dengan tepat dan 76 data yang salah diprediksi sebagai HS (*false positive*). Sementara itu, pada kelas HS, sebanyak 751 data berhasil diklasifikasikan dengan benar dan 84 data salah diprediksi sebagai NON-HS (*false negative*).

Jumlah *false negative* yang sedikit lebih tinggi dibandingkan *false positive* menunjukkan bahwa masih terdapat ujaran kebencian yang belum terdeteksi oleh model. Meskipun demikian, jumlah prediksi benar pada kedua kelas jauh lebih

besar dibandingkan kesalahannya, sehingga model tetap menunjukkan performa yang baik dan stabil.

Selanjutnya, pada Tahap 2 dilakukan klasifikasi kategori ujaran kebencian menggunakan pendekatan klasifikasi multi-label dengan model IndoBERTweet yang telah di *fine-tuning*. Model dikonfigurasi dengan *maximum sequence length* sebesar 512, *threshold* 0,5, serta dilatih selama 5 *epoch* dengan *batch size* 16 dan *learning rate* 5e-5. Data yang digunakan pada tahap ini berjumlah 5.553 data yang telah teridentifikasi sebagai ujaran kebencian (HS) pada Tahap 1. Berikut Tabel 4.16 yang berisi rekapitulasi dari 5 kali percobaan untuk tahap 2 dengan *seed* yang berbeda-beda. Dengan *seed* 42 sebagai *baseline* utama untuk analisis selanjutnya.

Tabel 4.16 Rekapitulasi Hasil Percobaan Tahap 2 Pada Skenario 2

Percobaan ke-	Macro Precision %	Macro Recall %	Macro F1-Score %	Hamming Loss %
1	76	82	78,63	9,39
2	77	84	80,51	9,27
3	78	83	80,13	8,51
4	75	81	78,71	9,99
5	75	81	77,73	9,52
Rata-Rata (Mean)	76	82	79	9
Standar Deviasi (Std)	1,30	1,30	1,14	0,54

Berdasarkan Tabel 4.16, hasil percobaan pada Tahap 2 dengan menggunakan teks augmentasi berbasis *Masked Language Model* (MLM) menunjukkan bahwa performa model cenderung stabil dengan nilai rata-rata (*mean*) *Macro F1-Score* sebesar 79% dan standar deviasi sebesar 1,14%. Nilai ini menunjukkan bahwa variasi performa antar percobaan masih tergolong rendah, meskipun sedikit lebih tinggi dibandingkan skenario tanpa augmentasi.

Dari sisi *precision* dan *recall*, model menunjukkan kecenderungan nilai *recall* 82% yang lebih tinggi dibandingkan *precision* 76%. Hal ini mengindikasikan bahwa penggunaan augmentasi MLM membantu model dalam mengenali lebih banyak variasi pola ujaran kebencian, sehingga meningkatkan kemampuan dalam mendeteksi label yang relevan. Namun, peningkatan *recall* ini diikuti dengan penurunan *precision*, yang menunjukkan bahwa model menjadi lebih permisif dan cenderung menghasilkan prediksi label yang lebih banyak, termasuk yang tidak tepat.

Selain itu, nilai *Hamming Loss* yang relatif rendah sekitar 9% menunjukkan bahwa kesalahan prediksi multi-label masih dapat dikendalikan dengan baik. Namun, standar deviasi *Hamming Loss* sebesar 0,54% yang lebih tinggi dibandingkan skenario sebelumnya menunjukkan bahwa penggunaan augmentasi juga dapat menambah variasi dalam kesalahan prediksi.

Secara keseluruhan, hasil ini menunjukkan bahwa penggunaan augmentasi teks berbasis MLM memberikan dampak positif dalam meningkatkan kemampuan model dalam menangkap variasi data ditunjukkan oleh peningkatan *recall*, namun dengan konsekuensi adanya penurunan *precision*. Hal ini mencerminkan adanya *trade-off* antara *recall* dan *precision* pada skenario dengan augmentasi.

Dengan demikian, meskipun model tetap menunjukkan kestabilan yang cukup baik terhadap variasi random seed, penggunaan augmentasi MLM perlu dipertimbangkan secara hati-hati karena dapat memengaruhi keseimbangan performa model dalam klasifikasi multi-label.

Untuk memahami lebih lanjut faktor yang memengaruhi performa tersebut, khususnya terkait perbedaan nilai precision dan recall, dilakukan analisis terhadap distribusi label pada dataset. Analisis ini bertujuan untuk mengidentifikasi adanya ketidakseimbangan kelas yang dapat memengaruhi kinerja model.

Hasil analisis distribusi label menunjukkan adanya ketidakseimbangan kelas, yang ditandai dengan perbedaan jumlah data antar kategori, di mana beberapa label memiliki jumlah sampel yang jauh lebih sedikit dibandingkan label lainnya. Oleh karena itu, untuk mengatasi ketidakseimbangan tersebut, dilakukan augmentasi data menggunakan pendekatan Masked Language Modeling (MLM) dengan menambahkan 60% data pada label minoritas. Distribusi jumlah data pada setiap kategori ujaran kebencian disajikan pada Tabel 4.17.

Tabel 4.17 Distribusi Data Pada Klasifikasi Multi-label Pada Skenario 2

Rekapitulasi jumlah Distribusi Data Pada Setiap Label							
	HS_ <i>Religion</i>	HS_ <i>Race</i>	HS_ <i>Individual</i>	HS_ <i>Group</i>	HS_ <i>Physical</i>	HS_ <i>Gender</i>	HS_ <i>Other</i>
<i>Training before MLM</i>	554	391	2484	1400	219	223	2616
<i>Training after MLM</i>	886	625	2867	1849	369	374	2616
<i>Validation</i>	100	80	489	260	44	43	505
<i>Testing</i>	105	77	487	264	54	34	503

Berdasarkan Tabel 4.17, distribusi data pada setiap label masih menunjukkan kondisi ketidakseimbangan (*imbalanced data*). Pada data training sebelum augmentasi, beberapa kategori memiliki jumlah data yang relatif rendah, seperti HS_*Physical* sebanyak 219 data, HS_*Gender* sebanyak 223 data, HS_*Race* sebanyak 391 data, dan HS_*Religion* sebanyak 554 data. Di sisi lain, kategori seperti HS_*Individual* (2.484 data), HS_*Group* (1.400 data), dan HS_*Other* (2.616 data) memiliki jumlah data yang jauh lebih besar. Ketimpangan distribusi ini dapat

menyebabkan model lebih dominan mempelajari pola pada kelas mayoritas sehingga performa klasifikasi pada label minoritas berpotensi kurang optimal.

Untuk mengatasi permasalahan tersebut, dilakukan augmentasi data menggunakan pendekatan *Masked Language Modeling* (MLM) berbasis model IndoBERTweet. Teknik ini dilakukan dengan melakukan masking pada sebagian token di dalam kalimat, kemudian model memprediksi token pengganti yang masih relevan secara kontekstual sehingga menghasilkan variasi teks baru tanpa mengubah makna utama kalimat. Proses augmentasi diterapkan pada label minoritas yang memiliki jumlah data training kurang dari 1.000, yaitu *HS_Religion*, *HS_Race*, *HS_Physical*, dan *HS_Gender*, dengan target penambahan sebesar 60% dari jumlah data awal.

Hasil augmentasi menunjukkan adanya peningkatan jumlah data pada label minoritas. *HS_Religion* meningkat dari 554 menjadi 886 data atau bertambah sebanyak 332 data. *HS_Race* meningkat dari 391 menjadi 625 data dengan penambahan 234 data. *HS_Physical* meningkat dari 219 menjadi 369 data atau bertambah 150 data, sedangkan *HS_Gender* meningkat dari 223 menjadi 374 data dengan tambahan 151 data. Peningkatan tersebut menunjukkan bahwa augmentasi berhasil memperbanyak representasi data pada kategori yang sebelumnya memiliki jumlah rendah sehingga distribusi data menjadi lebih seimbang.

Selain label minoritas, beberapa label dengan jumlah data yang sebelumnya sudah tinggi juga mengalami peningkatan meskipun tidak secara langsung menjadi target augmentasi. *HS_Individual* meningkat dari 2.484 menjadi 2.867 data atau bertambah 383 data, sedangkan *HS_Group* meningkat dari 1.400 menjadi 1.849

data atau bertambah 449 data. Hal ini terjadi karena karakteristik klasifikasi multi-label memungkinkan satu tweet memiliki lebih dari satu label secara bersamaan (label *co-occurrence*). Sebagai contoh, sebuah tweet dapat memiliki kombinasi label seperti *HS_Religion* dan *HS_Individual*, atau *HS_Race* dan *HS_Group* secara simultan. Ketika tweet pada label minoritas di-augment, label lain yang melekat pada tweet tersebut ikut terbawa ke data hasil augmentasi sehingga jumlahnya juga meningkat. Oleh karena itu, walaupun *HS_Individual* dan *HS_Group* tidak termasuk kategori yang secara eksplisit dilakukan augmentasi, jumlah datanya tetap bertambah sebagai konsekuensi dari hubungan antarlabel pada skenario multi-label.

Sementara itu, *HS_Other* tetap berada pada jumlah 2.616 data tanpa perubahan. Hal ini menunjukkan bahwa label tersebut tidak banyak muncul bersamaan dengan label minoritas yang menjadi target augmentasi, sehingga proses penambahan data tidak memberikan dampak signifikan pada kategori tersebut.

Secara keseluruhan, augmentasi MLM berhasil meningkatkan jumlah data training, terutama pada kategori minoritas, sekaligus mempertahankan hubungan alami antarlabel pada data multi-label. Proses augmentasi hanya diterapkan pada data training, sedangkan data *validation* dan *testing* tetap dipertahankan tanpa perubahan untuk menghindari data leakage dan menjaga validitas evaluasi model.

Berikut rekapitulasi jumlah label dalam 1 tweet pada Tabel 4.18

Tabel 4.18 Rekapitulasi Jumlah Label Dalam 1 Tweet Pada Skenario 2

Rekapitulasi Jumlah Label Dalam 1 Tweet							
	1	2	3	4	5	6	7
<i>Training</i>	-	3771	109	4	-	-	-
<i>Validation</i>	-	728	19	2	-	-	-
<i>Testing</i>	-	730	20	1	-	-	-

Berdasarkan Tabel 4.18 rekapitulasi jumlah label dalam satu tweet, terlihat bahwa sebagian besar data memiliki dua label dalam satu tweet, baik pada data training, validation, maupun testing. Hal ini menunjukkan bahwa karakteristik dataset bersifat multi-label dengan kecenderungan kombinasi label yang cukup kompleks, sehingga menambah tingkat kesulitan dalam proses klasifikasi.

Mengingat tugas klasifikasi dilakukan pada tujuh kategori ujaran kebencian dengan karakteristik linguistik yang beragam serta distribusi data yang tidak seimbang, nilai tersebut menunjukkan bahwa model telah mampu mempelajari sebagian besar pola label dengan cukup baik. Hasil evaluasi performa model untuk setiap kategori ditampilkan pada Tabel 4.19.

Tabel 4.19 Hasil Evaluasi Klasifikasi Multi-Label (Tahap 2) Pada Skenario 2

Label	<i>Precision</i> (%)	<i>Recall</i> (%)	<i>F1-Score</i> (%)
HS_Religion	70	86	77
HS_Race	74	87	80
HS_Individual	80	91	85
HS_Group	73	78	75
HS_Physical	80	69	74
HS_Gender	71	71	71
HS_Other	84	92	88
Macro Avg (%)	76	82	78,68
Hamming Loss (%)	9,39		

Berdasarkan Tabel 4.19, performa model pada setiap kategori menunjukkan hasil yang cukup baik dengan variasi nilai *precision*, *recall*, dan *F1-Score*. Secara umum, model mampu mengklasifikasikan sebagian besar kategori dengan cukup seimbang, meskipun masih terdapat perbedaan performa antar label.

Kategori HS_Other dan HS_Individual menunjukkan performa terbaik dengan *F1-Score* masing-masing sebesar 88% dan 85%. Hal ini mengindikasikan bahwa model mampu mengenali pola pada kategori dengan jumlah data yang lebih

besar secara lebih optimal. Sementara itu, kategori *HS_Race* dan *HS_Religion* juga menunjukkan performa yang baik, yang mengindikasikan bahwa proses augmentasi data membantu meningkatkan kemampuan model dalam mengenali kategori yang sebelumnya tergolong minoritas.

Di sisi lain, kategori *HS_Physical* dan *HS_Gender* memiliki performa yang relatif lebih rendah, dengan *F1-Score* sebesar 74% dan 71%. Hal ini menunjukkan bahwa model masih mengalami kesulitan dalam menangkap pola pada kategori dengan karakteristik yang lebih kompleks atau jumlah data yang lebih terbatas.

Secara keseluruhan, model memperoleh nilai *Macro F1-Score* sebesar 78,68%, yang menunjukkan performa yang cukup baik dalam menangani seluruh kategori secara seimbang. Nilai *Hamming Loss* sebesar 9,39% mengindikasikan bahwa tingkat kesalahan prediksi relatif rendah untuk tugas klasifikasi multi-label. Untuk analisis yang lebih rinci mengenai pola kesalahan prediksi, Tabel 4.20 menyajikan *confusion matrix* pada setiap kategori.

Tabel 4.20 Confusion Matrix Multi-label Pada Tahap 2 Skenario 2

label	TP	FP	FN	TN
<i>HS_Religion</i>	95	35	11	679
<i>HS_Race</i>	66	28	11	715
<i>HS_Individual</i>	432	112	51	225
<i>HS_Group</i>	200	76	62	482
<i>HS_Physical</i>	37	13	16	754
<i>HS_Gender</i>	29	14	6	772
<i>HS_Other</i>	449	75	47	249

Berdasarkan Tabel 4.20, performa model pada setiap label menunjukkan kemampuan yang berbeda dalam mengenali kategori ujaran kebencian. Kategori *HS_Other* dan *HS_Individual* memiliki jumlah *true positive* tertinggi, yaitu masing-masing sebesar 449 dan 432. Hal ini menunjukkan bahwa model mampu mengenali

sebagian besar data positif pada kedua kategori tersebut. Tingginya nilai tersebut kemungkinan dipengaruhi oleh jumlah data pelatihan yang lebih besar sehingga model lebih sering mempelajari pola bahasa yang merepresentasikan kedua label tersebut.

Selain itu, penerapan augmentasi berbasis *Masked Language Modeling* (MLM) turut membantu meningkatkan kemampuan model dalam memahami variasi konteks bahasa. Melalui proses masking dan prediksi token, model memperoleh representasi semantik yang lebih kaya sehingga mampu melakukan generalisasi dengan lebih baik pada data uji.

Pada kategori HS_Group, model juga menunjukkan kemampuan deteksi yang cukup baik dengan *true positive* sebesar 200. Namun, nilai *false positive* (76) dan *false negative* (62) yang relatif tinggi mengindikasikan bahwa model masih mengalami kesulitan membedakan ujaran kebencian terhadap kelompok dengan kategori lain yang memiliki konteks serupa. Kondisi ini menunjukkan adanya ambiguitas konteks pada ujaran yang mengarah pada identitas kelompok tertentu.

Untuk kategori HS_Religion dan HS_Race, model menunjukkan kemampuan klasifikasi yang cukup baik dengan *true positive* masing-masing sebesar 95 dan 66 serta nilai *false negative* yang relatif rendah, yaitu sama-sama 11. Hal ini mengindikasikan bahwa sebagian besar ujaran kebencian pada kategori tersebut berhasil dikenali, meskipun masih ditemukan *false positive* sebesar 35 pada HS_Religion dan 28 pada HS_Race yang menunjukkan adanya prediksi berlebih terhadap label tersebut.

Pada kategori HS_Physical dan HS_Gender, jumlah *true positive* relatif lebih rendah, yaitu masing-masing sebesar 37 dan 29. Hal ini dapat disebabkan oleh jumlah data positif yang lebih sedikit sehingga model memperoleh representasi pola yang lebih terbatas selama pelatihan. Meski demikian, kedua kategori memiliki nilai *false positive* dan *false negative* yang relatif rendah, yang menunjukkan bahwa kesalahan prediksi masih berada pada tingkat yang lebih terkendali dibandingkan beberapa kategori lain.

Secara keseluruhan, model menunjukkan kemampuan yang baik dalam mengenali kategori dengan distribusi data yang lebih besar, seperti HS_Individual dan HS_Other, namun masih menghadapi tantangan pada kategori dengan konteks yang saling tumpang tindih atau jumlah data yang lebih terbatas. Nilai *false positive* dan *false negative* pada beberapa label menunjukkan bahwa model masih mengalami kesulitan dalam membedakan karakteristik antar kategori ujaran kebencian yang memiliki kemiripan konteks bahasa.

Selanjutnya, pada Tahap 3 dilakukan klasifikasi tingkat keparahan (*severity*) ujaran kebencian ke dalam tiga kelas, yaitu *Weak*, *Moderate*, dan *Strong*. Distribusi data menunjukkan adanya ketidakseimbangan kelas, dengan kategori *Weak* sebanyak 3.231 data, *Moderate* sebanyak 1.641 data, dan *Strong* sebanyak 471 data, sehingga kelas mayoritas didominasi oleh kategori *Weak*. Ketidakseimbangan ini berpotensi mempengaruhi proses pembelajaran model karena model cenderung lebih banyak mempelajari pola pada kelas dengan jumlah data yang lebih besar.

Selanjutnya pada tahap 3 dilakukan klasifikasi tingkat keparahan (*severity*) ujaran kebencian ke dalam tiga kelas, yaitu *Weak*, *Moderate*, dan *Strong*. Distribusi

data menunjukkan adanya ketidakseimbangan kelas, dengan kategori *Weak* sebanyak 3.231 data, *Moderate* sebanyak 1.641 data, dan *Strong* sebanyak 471 data, sehingga kelas mayoritas didominasi oleh kategori *Weak*. Ketidakseimbangan ini berpotensi mempengaruhi proses pembelajaran model karena model cenderung lebih sering mempelajari pola pada kelas yang memiliki jumlah data lebih besar.

Pada tahap ini menggunakan beberapa *hyperparameter*, yaitu *max length* 512, *threshold* 0,5, *epoch* 5, *batch size* 16, dan *learning rate* 5e-5, dengan *backbone* *indolem/IndoBERTweet-base-uncased*, dengan menggunakan pendekatan CORN (*Conditional Ordinal Regression for Neural networks*) untuk menangani permasalahan klasifikasi ordinal. Berikut Tabel 4.21 yang berisi rekapitulasi dari beberapa percobaan yang sama dengan tahap sebelumnya.

Tabel 4.21 Rekapitulasi Hasil Percobaan Tahap 3 Pada Skenario 2

Percobaan ke-	Macro Precision %	Macro Recall %	Macro F1-Score %	Accuracy	QWK %
1	80,82	81,91	81,34	82,56	76,80
2	79,59	78,76	79,16	81,71	73,81
3	86,46	85,14	85,74	86,42	81,13
4	81,49	79,68	80,52	82,02	74,57
5	83,31	80,06	81,55	82,72	74,84
Rata-Rata (Mean)	82,33	81,11	81,66	83,09	76,23
Standar Deviasi (Std)	2,67	2,53	2,47	1,91	2,95

Berdasarkan Tabel 4.21 yang berisi hasil percobaan pada Tahap 3, model memperoleh rata-rata (*mean*) *macro F1-Score* sebesar 81,66% dengan standar deviasi sebesar 2,47. Nilai ini menunjukkan bahwa model memiliki performa yang cukup baik dalam melakukan klasifikasi tingkat keparahan, namun dengan tingkat variasi yang relatif tinggi antar percobaan.

Variasi performa juga terlihat pada nilai *macro precision* dan *macro recall* yang masing-masing memiliki standar deviasi sebesar 2,67% dan 2,53%. Hal ini mengindikasikan bahwa model cukup sensitif terhadap perubahan inisialisasi acak pada tahap ini. Sensitivitas ini dapat disebabkan oleh kompleksitas tugas klasifikasi ordinal, di mana model tidak hanya membedakan kelas, tetapi juga harus memahami hubungan tingkatan antar label.

Selain itu, nilai QWK sebagai metrik evaluasi ordinal memiliki rata-rata sebesar 76,23% dengan standar deviasi sebesar 2,95%, yang merupakan nilai variasi tertinggi dibandingkan metrik lainnya. Hal ini menunjukkan bahwa konsistensi model dalam menangkap hubungan urutan tingkat keparahan masih belum stabil. Dengan kata lain, model belum sepenuhnya mampu mempertahankan konsistensi dalam memprediksi kedekatan antar level, seperti perbedaan antara kategori *weak*, *moderate*, dan *strong*.

Perbedaan performa yang cukup mencolok pada percobaan ke3 dibandingkan percobaan lainnya menunjukkan bahwa konfigurasi tertentu dapat menghasilkan performa yang lebih optimal. Hal ini mengindikasikan bahwa model memiliki potensi performa yang tinggi, namun belum konsisten pada seluruh percobaan.

Secara keseluruhan, hasil ini menunjukkan bahwa Tahap 3 merupakan tahap yang paling menantang dalam *pipeline* klasifikasi hierarkis. Meskipun model mampu mencapai performa yang baik, tingkat kestabilannya lebih rendah dibandingkan tahap sebelumnya. Kondisi ini menunjukkan bahwa klasifikasi tingkat keparahan menjadi faktor kritis yang dapat memengaruhi performa akhir

sistem secara keseluruhan. Untuk hasil evaluasi pada setiap level keparahan selanjutnya disajikan pada Tabel 4.22.

Tabel 4.22 Hasil Evaluasi Tingkat Keparahan Pada Tahap 3 Skenario 2

Severity	Precision (%)	Recall (%)	F1-Score (%)
Weak	87,99	87,47	87,28
Moderate	73,42	71,49	72,44
Strong	81,94	86,76	84,29
Accuracy (%)	82,56		
Quadratic Weighted Kappa (QWK) (%)	76,80		

Berdasarkan hasil evaluasi pada Tabel 4.22, model menunjukkan kinerja yang cukup baik dalam mengklasifikasikan tingkat keparahan ujaran kebencian. Kategori *Weak* memperoleh nilai *F1-Score* tertinggi sebesar 87,28% dengan *recall* sebesar 87,47%, yang menunjukkan bahwa model mampu mengenali sebagian besar data pada kategori ini secara konsisten. Hal ini wajar mengingat kategori *Weak* merupakan kelas dengan jumlah data terbanyak, sehingga pola yang dipelajari model menjadi lebih kuat.

Kategori *Strong* menunjukkan karakteristik yang menarik, dengan nilai *recall* yang tinggi yaitu 86,76% namun *precision* yang lebih rendah yaitu 81,94%. Kondisi ini mengindikasikan bahwa model cukup sensitif dalam mendeteksi ujaran dengan tingkat keparahan tinggi, tetapi masih menghasilkan cukup banyak *false positive*. Artinya, beberapa data dari kategori lain terutama *moderate* cenderung diprediksi sebagai *strong*, yang menunjukkan adanya kecenderungan model untuk *over-predict* pada kelas dengan intensitas tinggi.

Sementara itu, kategori *moderate* memiliki performa terendah dengan *F1-Score* sebesar 72,44%, serta *recall* yang relatif lebih rendah 73,24%. Hal ini

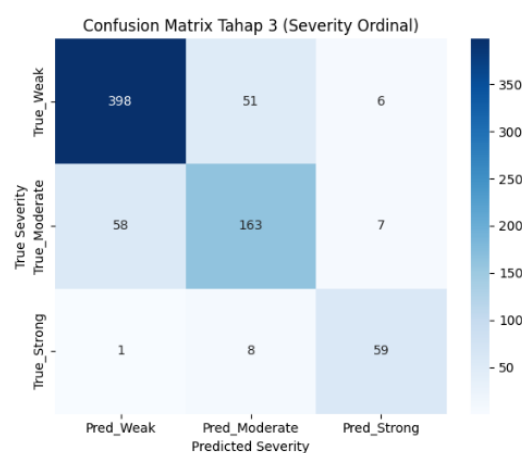
menunjukkan bahwa model masih kesulitan dalam mengenali kelas *moderate* secara konsisten. Secara karakteristik, kelas *moderate* memang berada di antara dua kelas lain *weak* dan *strong*, sehingga memiliki batas yang lebih ambigu dan sering mengalami tumpang tindih secara semantik.

Kinerja model ini juga dipengaruhi oleh penggunaan pendekatan *Conditional Ordinal Regression for Neural Networks* (CORN), yang memodelkan klasifikasi sebagai permasalahan ordinal (berurutan), bukan sekadar klasifikasi biasa. Dalam CORN, prediksi dilakukan secara bertahap melalui beberapa *threshold*, misalnya memprediksi apakah suatu data melewati batas *weak* ke *moderate*, kemudian *moderate* ke *strong*. Mekanisme ini membuat model lebih memahami hubungan tingkat keparahan antar kelas.

Dampak dari pendekatan CORN terlihat pada pola kesalahan model yang cenderung terjadi pada kelas yang berdekatan. Hal ini tercermin dari nilai *Quadratic Weighted Kappa* (QWK) sebesar 76,80%, yang menunjukkan bahwa prediksi model memiliki tingkat kesepakatan yang baik dengan label sebenarnya dalam konteks ordinal. Selain itu, pendekatan ini juga menyebabkan model menjadi lebih sensitif terhadap kelas dengan tingkat keparahan tinggi, sehingga recall pada kategori Strong menjadi tinggi, meskipun diikuti dengan peningkatan *false positive*.

Namun, pendekatan CORN juga memiliki keterbatasan, terutama pada kelas tengah seperti *moderate*. Karena posisi *moderate* berada di antara dua batas keputusan, model harus menentukan apakah suatu data lebih dekat ke *weak* atau *strong*. Hal ini menyebabkan kelas *moderate* menjadi yang paling sulit diprediksi dan sering mengalami pergeseran ke kelas tetangga.

Secara keseluruhan, model memperoleh akurasi sebesar 82,56%, yang menunjukkan performa yang baik. Perbedaan performa antar kategori dipengaruhi oleh ketidakseimbangan distribusi data serta kedekatan semantik antar kelas. Untuk melihat pola kesalahan secara lebih rinci, Gambar 4.5 menyajikan *confusion matrix* pada Tahap 3, yang dapat digunakan untuk mengidentifikasi pergeseran prediksi antar tingkat keparahan.



Gambar 4.5 *Confusion Matrix* Tingkat Keparahan Tahap 3 Skenario 2

Berdasarkan Gambar 4.5, terlihat bahwa pola kesalahan prediksi model didominasi oleh pergeseran antar kelas yang berdekatan. Kesalahan paling banyak terjadi pada perpindahan antara *moderate* ke *weak* (58 data) dan *weak* ke *moderate* (51 data), yang menunjukkan bahwa batas antara kedua kelas tersebut masih belum terdefinisi dengan jelas oleh model.

Selain itu, jumlah kesalahan langsung antara kelas yang berjauhan relatif sangat kecil, seperti dari *strong* ke *weak* (1 data) dan *weak* ke *strong* (6 data). Pola ini menunjukkan bahwa model jarang melakukan kesalahan ekstrem, sehingga prediksi yang dihasilkan masih mengikuti urutan tingkat keparahan secara logis.

Fenomena ini mengindikasikan bahwa model lebih sering mengalami ketidakpastian dalam menentukan intensitas ujaran pada tingkat menengah, dibandingkan kesalahan dalam mengenali perbedaan ekstrem. Dengan kata lain, tantangan utama model terletak pada penentuan batas antar kelas yang berdekatan, bukan pada pemahaman perbedaan tingkat keparahan secara keseluruhan.

Pola kesalahan yang bersifat lokal ini juga mencerminkan dampak dari pendekatan CORN, di mana proses klasifikasi dilakukan secara bertahap berdasarkan batas antar level. Akibatnya, ketika terjadi kesalahan, prediksi cenderung bergeser ke kelas terdekat, bukan langsung ke kelas yang jauh.

Selanjutnya dilakukan evaluasi *end-to-end* (E2E) pada skenario 2 untuk menilai kinerja sistem secara keseluruhan dengan mempertimbangkan keterkaitan antar Tahap 1, Tahap 2, dan Tahap 3. Pada evaluasi ini, suatu prediksi dinyatakan benar hanya apabila seluruh tahapan klasifikasi berhasil diprediksi secara konsisten. Digunakan dua pendekatan evaluasi, yaitu *E2E HS Only* dan *E2E System*. Berikut Tabel 4.23 yang berisi rekapitulasi akurasi E2E dari 5 kali percobaan dengan *seed* 42 sebagai *baseline* utama penelitian ini.

Tabel 4.23 Rekapitulasi Hasil Percobaan E2E Pada Skenario 2

Percobaan ke-	E2E HS %	E2E System %
1	76,43	80,33
2	75,74	79,75
3	77,82	80,80
4	73,76	79,05
5	74,80	79,57
Rata-Rata (Mean)	75,71	79,90
Standar Deviasi (Std)	1,55	0,68

Berdasarkan Tabel 4.23, hasil evaluasi *end-to-end* (E2E) pada skenario 2 menunjukkan bahwa performa sistem secara keseluruhan relatif stabil. Hal ini terlihat dari nilai rata-rata (*mean*) *E2E System* sebesar 79,90% dengan standar deviasi sebesar 68%, yang mengindikasikan bahwa variasi performa antar percobaan tergolong rendah.

Sementara itu, nilai E2E HS memiliki rata-rata sebesar 75,71% dengan standar deviasi sebesar 1,55%, yang menunjukkan variasi yang lebih tinggi dibandingkan *E2E System*. Hal ini mengindikasikan bahwa performa sistem dalam menangani data ujaran kebencian (HS) lebih sensitif terhadap perubahan random seed, terutama karena data tersebut harus melalui seluruh tahapan klasifikasi secara berurutan.

Perbedaan nilai antara *E2E System* dan E2E HS menunjukkan adanya dampak dari proses klasifikasi hierarkis, di mana kesalahan pada Tahap 1 dapat menyebabkan sebagian data HS tidak diteruskan ke tahap berikutnya. Selain itu, kesalahan pada Tahap 2 dan Tahap 3 juga dapat terakumulasi, sehingga memengaruhi hasil akhir sistem. Kondisi ini menunjukkan adanya efek propagasi error dalam pipeline klasifikasi.

Jika dikaitkan dengan penggunaan augmentasi MLM pada skenario ini, variasi yang lebih tinggi pada E2E HS dapat disebabkan oleh peningkatan kompleksitas data hasil augmentasi, yang meskipun membantu model dalam menangkap variasi pola, juga berpotensi menambah noise. Hal ini berdampak pada ketidakkonsistenan performa pada beberapa percobaan.

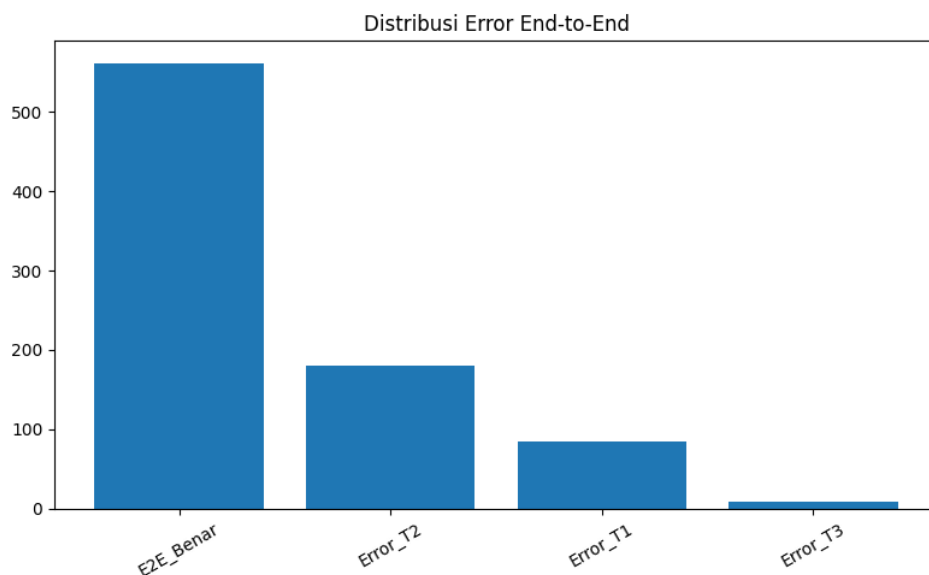
Meskipun demikian, nilai standar deviasi yang relatif kecil pada *E2E System* menunjukkan bahwa secara keseluruhan sistem tetap memiliki kestabilan yang baik. Dengan demikian, dapat disimpulkan bahwa penggunaan augmentasi MLM tidak secara signifikan mengganggu kestabilan sistem, namun memberikan pengaruh terhadap variasi performa khususnya pada data ujaran kebencian. Ringkasan hasil evaluasi E2E pada skenario 2 dengan *seed* 42 disajikan pada Tabel 4.24 yang menampilkan distribusi prediksi benar serta kesalahan pada setiap tahap.

Tabel 4.24 Hasil dan distribusi kesalahan pada evaluasi E2E skenario 2

<i>E2E HS Only</i>		<i>E2E System</i>	
Jumlah dataset	1.714	Jumlah test	1.714
Jumlah benar terdeteksi HS	751	Jumlah benar NON-HS	803
Jumlah benar tahap 1	751	Jumlah benar tahap 1	751
Jumlah benar tahap 2	584	Jumlah benar tahap 2	584
Jumlah benar tahap 3	574	Jumlah benar tahap 3	574
E2E Akurasi	76,43%	E2E Akurasi	80,33%

Berdasarkan Tabel 4.24 evaluasi *sistem end-to-end* (E2E), pengujian dilakukan pada dua skenario berbeda, yaitu *E2E HS Only* dan *E2E System*, dengan jumlah data testing sebanyak 1.714 tweet. Pada skenario *E2E HS Only*, evaluasi hanya memperhitungkan data yang memang berlabel ujaran kebencian, di mana 751 tweet berhasil diprediksi benar sebagai HS. Dari data ini, Tahap 2 (multi-label kategori HS) memprediksi 584 tweet dengan benar, sedangkan Tahap 3 (ordinal severity) sebanyak 574 tweet, sehingga diperoleh E2E Accuracy sebesar 76,43%. Sementara itu, pada skenario *E2E System*, evaluasi mengikuti alur pipeline sistem, di mana 803 tweet NON-HS otomatis dianggap benar. Dari tweet yang melewati Tahap 1, Tahap 2 memprediksi 584 tweet dengan benar dan Tahap 3 sebanyak 574 tweet, sehingga diperoleh *E2E Accuracy* sebesar 80,33%. Distribusi kesalahan menunjukkan bahwa Tahap 2 masih menjadi *bottleneck* utama dengan 167 tweet

salah prediksi pada kedua skenario, sedangkan Tahap 3 relatif stabil dengan hanya 10 kesalahan. Kesalahan pada Tahap 1 sebanyak 84 menunjukkan bahwa kesalahan awal dapat mempengaruhi performa sistem *end-to-end* secara signifikan. Secara keseluruhan, performa sistem lebih dipengaruhi oleh Tahap 2, sementara mekanisme *filtering* di Tahap 1 berperan penting untuk menjaga akurasi sistem. Untuk analisis lebih lanjut, Gambar 4.6 menyajikan visualisasi distribusi kesalahan pada evaluasi *end-to-end* skenario ini.



Gambar 4.6 Distribusi data benar E2E skenario 2

Gambar 4.6 menampilkan visualisasi distribusi hasil evaluasi end-to-end pada skenario 2 menggunakan dua pendekatan sebagian besar data berhasil diprediksi dengan benar secara keseluruhan oleh sistem. Kesalahan yang terjadi pada pendekatan ini terutama berasal dari Tahap 2, yaitu tahap klasifikasi kategori ujaran kebencian. Hal ini menunjukkan bahwa meskipun data telah berhasil melewati tahap deteksi awal, proses penentuan kategori ujaran kebencian masih menjadi tantangan utama bagi model.

Secara keseluruhan, hasil evaluasi menunjukkan bahwa performa sistem klasifikasi ujaran kebencian dipengaruhi oleh interaksi antar tahapan dalam pipeline. Tahap 2 menjadi bottleneck utama karena kompleksitas klasifikasi multi-label, sementara Tahap 3 relatif stabil dengan kontribusi kesalahan yang minimal. Perbedaan performa antara pendekatan *HS Only* dan *System* menegaskan bahwa kesalahan pada Tahap 1 memiliki dampak signifikan terhadap akurasi *end-to-end*. Dengan demikian, peningkatan kinerja sistem secara keseluruhan perlu difokuskan pada perbaikan klasifikasi kategori pada Tahap 2 serta optimalisasi deteksi awal pada Tahap 1.

4.1.3 Hasil Uji Skenario 3

Pada skenario 3, dilakukan pengujian menggunakan pendekatan klasifikasi *flat* sebagai model pembanding (*baseline*). Model yang digunakan adalah *backbone indolem/IndoBERTweet-base-uncased* dalam melakukan klasifikasi multi-label ujaran kebencian secara langsung dalam satu tahap.

Berbeda dengan pendekatan hierarkis pada skenario sebelumnya, model *flat* tidak mempertimbangkan struktur bertingkat antar label, seperti pemisahan kategori ujaran kebencian maupun tingkat keparahan. Seluruh label dipelajari secara simultan dalam satu model klasifikasi. Oleh karena itu, pendekatan ini digunakan sebagai baseline untuk mengevaluasi kontribusi struktur hierarkis terhadap performa sistem. Penelitian ini melakukan beberapa percobaan dengan *seed* yang berbeda untuk melihat kestabilan performa model. Rekapitulasi hasil evaluasi ditunjukkan pada Tabel 4.25.

Tabel 4.25 Rekapitulasi Hasil Evaluasi Pada Skenario 3

Percobaan ke-	Macro Precision %	Macro Recall %	Macro F1-Score %	Subset Accuracy %
1	69,22	68,71	68,63	82,15
2	69,92	69,22	69,44	82,73
3	72,60	72,35	72,25	82,15
4	68,43	65,38	66,60	81,86
5	63,25	66,60	64,73	80,22
Rata-Rata (Mean)	68,68	68,45	68,34	81,82
Standar Deviasi (Std)	3,42	2,68	2,86	0,95

Berdasarkan hasil pengujian pada Tabel 4.25, diperoleh nilai rata-rata (*mean*) *Macro Precision* sebesar 68,68%, *Macro Recall* sebesar 68,45%, dan *Macro F1-Score* sebesar 68,34%, dengan *Subset Accuracy* sebesar 81,82%. Nilai standar deviasi yang relatif kecil menunjukkan bahwa performa model cukup stabil meskipun menggunakan variasi *seed* yang berbeda.

Nilai *Subset Accuracy* yang relatif tinggi tidak sepenuhnya mencerminkan kemampuan model dalam mendeteksi seluruh label positif, melainkan juga dipengaruhi oleh tingginya jumlah prediksi benar pada label negatif (*non-HS*). Pada evaluasi multi-label, suatu sampel dianggap benar hanya jika seluruh labelnya sesuai, sehingga prediksi yang benar pada label yang tidak aktif (non-ujaran kebencian) turut meningkatkan nilai ini. Kondisi ini membuat *Subset Accuracy* cenderung lebih optimistis dibandingkan metrik berbasis per-label seperti *Macro F1-Score*.

Hal ini berbeda dengan pendekatan hierarkis, di mana evaluasi dilakukan secara bertahap sehingga kesalahan pada satu level dapat lebih terkontrol. Sementara pada pendekatan *flat*, seluruh label diprediksi secara simultan sehingga

kontribusi dominan berasal dari kelas non-aktif, yang pada akhirnya dapat menyebabkan bias pada nilai *Subset Accuracy*. Berikut hasil evaluasi dari percobaan pada *seed* 42 dari salah satu percobaan disajikan pada Tabel 4.26.

Tabel 4.26 Metrik Evaluasi Pendekatan Flat Klasifikasi

<i>Macro Precision</i> %	<i>Macro Recall</i> %	<i>Macro F1-Score</i> %	<i>Subset Accuracy</i> %
69,22%	68,71%	68,68%	82,15%

Berdasarkan Tabel 4.26, pendekatan tersebut menghasilkan *Subset Accuracy* sebesar 82,15%, yang menunjukkan bahwa sebagian besar sampel memiliki prediksi label yang sepenuhnya sesuai dengan *ground truth*. Namun, nilai ini tidak hanya dipengaruhi oleh keberhasilan dalam mendeteksi label positif, tetapi juga oleh tingginya keberhasilan model dalam mengklasifikasikan label negatif (non-ujaran kebencian), yang secara jumlah mendominasi dataset multi-label. Hal ini menyebabkan *Subset Accuracy* cenderung lebih optimistis dibandingkan metrik berbasis per-label.

Nilai *Macro Precision* sebesar 69,22% menunjukkan bahwa sebagian besar prediksi positif yang dihasilkan model merupakan label yang benar, sedangkan *Macro Recall* sebesar 68,71% mengindikasikan masih adanya sejumlah label ujaran kebencian yang tidak berhasil terdeteksi. Kondisi ini menunjukkan bahwa model memiliki kecenderungan konservatif dalam memberikan prediksi positif, sehingga lebih menghindari *false positive* namun berdampak pada meningkatnya *false negative*.

Sementara itu, nilai *Macro F1-Score* sebesar 68,63% menunjukkan keseimbangan antara *Macro Precision* dan *Recall*, meskipun masih terdapat ketimpangan dalam kemampuan model menangkap seluruh label secara merata.

Untuk analisis yang lebih rinci pada setiap label, digunakan *confusion matrix* yang ditunjukkan pada Tabel 4.27.

Tabel 4.27 *Confusion Matrix* Hasil Klasifikasi Flat Per Label

Label	True Positive (TP)	False Positive (FP)	False Negative (FN)	True Negative (TN)
HS_Religion	71	27	40	1576
HS_Race	64	31	19	1600
HS_Individual	446	111	99	1058
HS_Group	208	64	74	1368
HS_Physical	22	4	16	1672
HS_Gender	25	12	27	1650
HS_Other	461	115	104	1034
HS_Weak	415	115	99	1085
HS_Moderate	162	67	78	1407
HS_Strong	59	13	14	1628

Berdasarkan Tabel 4.27, performa model menunjukkan variasi pada setiap label yang diuji. Label dengan jumlah data yang lebih besar, seperti HS_Individual dan HS_Other, memiliki nilai *true positive* yang relatif tinggi, masing-masing sebesar 446 dan 461. Hal ini mengindikasikan bahwa model lebih mampu mengenali pola pada kelas yang memiliki distribusi data dominan, sehingga proses pembelajaran pada kategori tersebut menjadi lebih optimal dibandingkan kelas lainnya.

Sebaliknya, pada label dengan jumlah data yang lebih sedikit seperti HS_Physical, HS_Gender, HS_Religion, dan HS_Race, nilai *true positive* cenderung lebih rendah meskipun *true negative* yang dihasilkan cukup tinggi. Kondisi ini menunjukkan bahwa model lebih sering berhasil mengklasifikasikan data non-positif dengan benar, namun masih mengalami kesulitan dalam mendeteksi kelas minoritas secara tepat. Hal ini mengindikasikan adanya pengaruh ketidakseimbangan data terhadap kemampuan model dalam mengenali seluruh kategori secara merata.

Selain itu, masih ditemukan *false negative* pada beberapa label utama seperti *HS_Individual*, *HS_Other*, dan *HS_Weak*. Hal ini menunjukkan bahwa sebagian data ujaran kebencian tidak berhasil teridentifikasi oleh model, sehingga berdampak pada penurunan nilai *recall* secara keseluruhan. Kondisi ini sejalan dengan hasil evaluasi pada Tabel 4.27 yang menunjukkan bahwa model belum optimal dalam menangkap seluruh data positif yang ada.

Secara keseluruhan, hasil evaluasi ini menunjukkan bahwa pendekatan flat mampu memberikan performa dasar yang cukup baik dalam tugas klasifikasi ujaran kebencian. Namun demikian, pendekatan ini masih memiliki keterbatasan, terutama dalam menangani ketidakseimbangan data serta kompleksitas hubungan antar label, jika dibandingkan dengan pendekatan hierarkis yang digunakan pada skenario sebelumnya.

4.2 Pembahasan

Berdasarkan uji coba yang sudah dijabarkan pada subbab 4.1 sampai 4.1.3, pada bagian ini menyajikan analisis menyeluruh yang bertujuan untuk mengetahui konfigurasi secara maksimal dan pengaruh dari karakteristik dataset terhadap performa model.

4.2.1 Analisis Performa Model Terbaik

Berdasarkan hasil pengujian pada Skenario 1, Skenario 2, dan Skenario 3, performa model menunjukkan adanya perbedaan karakteristik yang dipengaruhi oleh distribusi data serta pendekatan yang digunakan. Oleh karena itu, penentuan model terbaik tidak hanya didasarkan pada nilai metrik evaluasi secara umum,

tetapi juga mempertimbangkan kemampuan model dalam menangani ketidakseimbangan data dan kompleksitas tugas klasifikasi. Berikut Tabel 4.28 yang berisi rekapitulasi dari evaluasi pada 3 skenario

Tabel 4.28 Rekapitulasi Evaluasi Pada 3 Skenario

Tahap 1						
	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)		
Skenario 1	90,81%	89,94%	90,37%	90,67%		
Skenario 2	90,81%	89,94%	90,37%	90,67%		
Tahap 2						
	Macro Precision (%)	Macro Recall (%)	Macro F1-Score (%)	Hamming Loss (%)		
Skenario 1	76%	81%	78,41%	9,5%		
Skenario 2	76%	82%	78,68%	9,39%		
Tahap 3						
	Macro Precision (%)	Macro Recall (%)	Macro F1-Score (%)	Accuracy (%)	Quadratic Weighted Kappa (%)	
Skenario 1	79,06 %	82,28 %	80,40%	83,22%	75,57%	
Skenario 2	80,82%	81,91%	81,34%	82,56%	76,80%	
E2E						
	E2E HS Only (%)			E2E System (%)		
Skenario 1	74,53%			79,43%		
Skenario 2	76,43%			80,33%		
Skenario 3 (Flat)						
	Macro Precision (%)	Macro Recall (%)	Macro F1-Score (%)	Subset Accuracy (%)		
Skenario 3	69,22%	68,71%	68,68%	82,15%		

Berdasarkan Tabel 4.28, Skenario 1 yang menggunakan data asli tanpa penanganan *class imbalance* menunjukkan performa yang relatif stabil pada distribusi data yang tersedia. Model mampu mempelajari pola dominan dengan baik, khususnya pada tahap klasifikasi biner. Namun, pada tahap klasifikasi multi-label, performa model cenderung menurun, terutama dalam mendeteksi label dengan jumlah data yang lebih sedikit. Hal ini mengindikasikan bahwa model

cenderung lebih optimal dalam mengenali label dengan distribusi dominan dibandingkan label minoritas.

Skenario 2 menunjukkan peningkatan performa pada tahap klasifikasi multi-label dibandingkan Skenario 1. Penerapan teknik *oversampling* berbasis *text augmentation* mampu memperbaiki distribusi data, sehingga model memiliki kemampuan yang lebih seimbang dalam mengenali berbagai kategori ujaran kebencian. Peningkatan ini terlihat dari kenaikan nilai *Macro Recall* dari 81% menjadi 82%, *Macro F1-Score* dari 78,41% menjadi 78,68%, serta penurunan *Hamming Loss* dari 9,5% menjadi 9,39%. Meskipun demikian, hasil evaluasi menunjukkan bahwa tahap klasifikasi multi-label cenderung menjadi *bottleneck* utama dalam sistem karena kompleksitas tugas yang tinggi. Dengan kata lain, meskipun balancing data memberikan dampak positif, tantangan dalam klasifikasi multi-label belum sepenuhnya dapat diatasi.

Di sisi lain, pendekatan klasifikasi hierarkis yang digunakan pada Skenario 1 dan Skenario 2 memberikan keunggulan dalam memodelkan hubungan antar label secara bertingkat. Namun, pendekatan ini juga memiliki kelemahan berupa propagasi kesalahan antar tahap. Kesalahan pada Tahap 1 menyebabkan data tidak dapat diteruskan ke tahap berikutnya, sehingga berdampak langsung terhadap performa sistem secara keseluruhan. Sementara itu, Tahap 3 menunjukkan performa yang relatif stabil dengan kontribusi kesalahan yang minimal, sehingga bukan menjadi faktor utama penurunan kinerja sistem.

Selanjutnya, Skenario 3 menggunakan pendekatan klasifikasi flat tanpa mempertimbangkan struktur hierarkis antar label. Hasil evaluasi menunjukkan

bahwa model memperoleh *Macro F1-Score* sebesar 68,68% dengan *Subset Accuracy* sebesar 82,15%. Meskipun pendekatan ini mampu memberikan performa evaluasi yang cukup baik, model masih memiliki keterbatasan dalam memahami hubungan antar label secara bertingkat. Selain itu, nilai *recall* yang lebih rendah dibandingkan skenario hierarkis mengindikasikan bahwa model masih belum optimal dalam mendeteksi seluruh kasus ujaran kebencian, sehingga kemungkinan terjadinya *false negative* masih cukup tinggi pada beberapa kategori. Hal ini menunjukkan bahwa meskipun performa model secara umum terlihat cukup baik, pendekatan flat belum mampu menangkap kompleksitas struktur data secara menyeluruh.

Secara keseluruhan, Skenario 1 dan Skenario 2 menunjukkan performa yang relatif serupa pada sebagian besar tahap pengujian. Namun, Skenario 2 cenderung memberikan peningkatan performa pada tahap klasifikasi multi-label, yang terlihat dari kenaikan nilai *Macro Recall* dan *Macro F1-Score* serta penurunan *Hamming Loss* dibandingkan Skenario 1. Hal ini menunjukkan bahwa penerapan *balancing* data memberikan pengaruh positif terhadap kemampuan model dalam menangani ketidakseimbangan data. Di sisi lain, Skenario 1 tetap menunjukkan performa yang stabil pada distribusi data asli. Oleh karena itu, meskipun perbedaan performa antar skenario tidak terlalu besar, masing-masing pendekatan memiliki karakteristik dan keunggulan tersendiri dalam proses klasifikasi.

Meskipun terdapat perbedaan nilai evaluasi secara deskriptif, diperlukan pengujian statistik untuk mengetahui apakah perbedaan tersebut signifikan atau tidak. Berikut Tabel 4.29 yang berisi hasil uji t berpasangan (*paired t-test*) dengan

pendekatan *one-tailed* yang digunakan untuk menguji apakah terdapat perbedaan performa yang signifikan antara semua skenario berdasarkan metrik evaluasi pada setiap tahap pengujian.

Tabel 4.29 Tabel *t-test* Antara Skenario 1, 2, dan 3

	Perbandingan	Matrix	<i>t-value</i>	<i>p-value</i>	Signifikan
Tahap 1	S1 VS S2	<i>F1-Score</i>	0,0919	0,4655	Tidak
	S1 VS S2	<i>Accuracy</i>	-1,1097	0,1646	Tidak
Tahap 2	S1 VS S2	<i>Macro F-1 Score</i>	-0,9646	0,1946	Tidak
		<i>Hamming Loss</i>	0,4167	0,3491	Tidak
Tahap 3	S1 VS S2	<i>QWK</i>	-1,5021	0,1037	Tidak
Tahap 3	S1 VS S2	<i>Macro F1 Score</i>	-1,000	0,1869	Tidak
		<i>Accuracy</i>	1,3544	0,1235	Tidak
E2E	S1 VS S2	<i>E2E System</i>	-1,4314	0,1127	Tidak
	S1 VS S3	<i>E2E System VS Subset Accuracy</i>	-4,5477	0,0052	Signifikan
	S2 VS S3	<i>E2E System VS Subset Accuracy</i>	-4,372	0,0059	Signifikan

Berdasarkan Tabel 4.29, hasil uji *t* berpasangan (*paired t-test*) dengan pendekatan *one-tailed* menunjukkan bahwa sebagian besar perbandingan antara Skenario 1 dan Skenario 2 menghasilkan nilai *p-value* di atas tingkat signifikansi $\alpha = 0,05$. Hal ini menunjukkan bahwa penerapan *balancing* data pada Skenario 2 belum memberikan peningkatan performa yang signifikan secara statistik dibandingkan Skenario 1 pada seluruh tahap pengujian.

Pada Tahap 1, metrik *F1-score* dan *accuracy* masing-masing menghasilkan *p-value* sebesar 0,4655 dan 0,1646. Nilai ini menunjukkan bahwa tidak terdapat perbedaan signifikan dalam hal keseimbangan antara *precision* dan *recall* maupun akurasi keseluruhan antara Skenario 1 dan Skenario 2, karena nilai lebih dari $\alpha = 0,05$. Selain itu, nilai *t-value* yang mendekati nol pada *F1-Score* mengindikasikan bahwa selisih performa antar skenario sangat kecil, sehingga perubahan skenario tidak memberikan dampak berarti pada kemampuan klasifikasi biner.

Pada Tahap 2, *Macro F1-Score* menghasilkan *p-value* sebesar 0,1946, yang menunjukkan bahwa peningkatan performa rata-rata pada seluruh kelas tidak memberikan dampak signifikan. Sementara itu, *Hamming Loss* memiliki *p-value* sebesar 0,3491, yang mengindikasikan bahwa tingkat kesalahan prediksi label antar skenario juga tidak berbeda secara signifikan. Hal ini menunjukkan bahwa penerapan teknik *oversampling* belum mampu secara konsisten meningkatkan performa maupun menurunkan tingkat kesalahan pada klasifikasi multi-label.

Pada Tahap 3, metrik *Quadratic Weighted Kappa (QWK)* menghasilkan *p-value* sebesar 0,1037, yang relatif lebih mendekati ambang signifikansi dibandingkan metrik lainnya. Hal ini mengindikasikan adanya kecenderungan peningkatan dalam kesesuaian tingkat keparahan prediksi, meskipun belum signifikan secara statistik. Sementara itu, *Macro F1-Score* dan *accuracy* masing-masing memiliki *p-value* sebesar 0,1869 dan 0,1235, yang menunjukkan bahwa baik keseimbangan performa antar kelas maupun akurasi keseluruhan tidak mengalami peningkatan yang berarti antar skenario.

Pada evaluasi *end-to-end (E2E)*, perbandingan antar Skenario 1 dan Skenario 2 menunjukkan nilai *p-value* 0,1127 sehingga menunjukkan perbedaan yang tidak signifikan karena masih di atas $\alpha = 0,05$. Namun, perbandingan S1 vs S3 dan S2 vs S3 masing-masing menghasilkan *p-value* sebesar 0,0052 dan 0,0059. Nilai tersebut berada di bawah tingkat signifikansi $\alpha = 0,05$, sehingga menunjukkan adanya perbedaan performa yang signifikan antara pendekatan klasifikasi hierarkis dan *flat classification* pada evaluasi sistem secara keseluruhan.

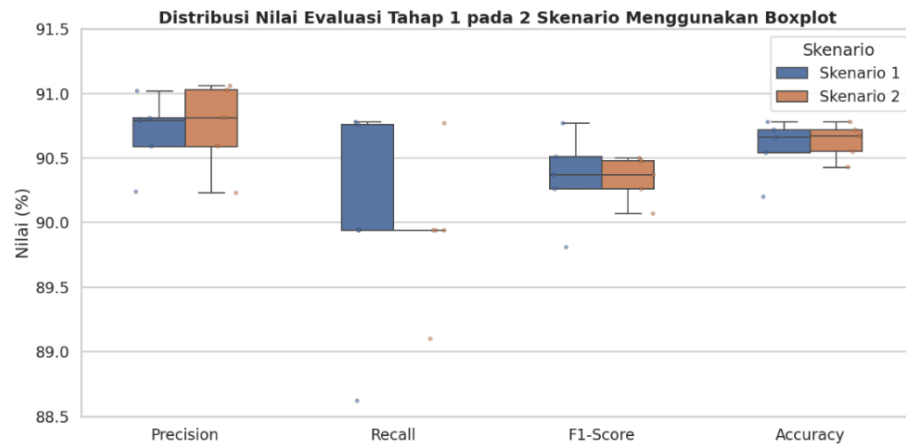
Tidak signifikannya Skenario 1 dan Skenario 2 juga dipengaruhi oleh beberapa faktor. Pertama desain eksperimen yang memiliki variasi terbatas. Perbedaan antar skenario hanya terletak pada penerapan teknik *balancing*, sementara aspek lain seperti arsitektur model, *hyperparameter*, dan *pipeline* pelatihan tetap sama. Akibatnya, ruang perubahan performa menjadi sempit, sehingga perbedaan yang dihasilkan cenderung kecil dan sulit terdeteksi secara statistik. Selain itu, jumlah percobaan yang terbatas (lima variasi *seed*) juga menyebabkan kekuatan uji statistik dalam mendeteksi perbedaan menjadi relatif rendah.

Meskipun demikian, secara deskriptif Skenario 2 tetap menunjukkan performa yang lebih konsisten, khususnya dalam menangani ketidakseimbangan data pada tahap multi-label. Oleh karena itu, pemilihan Skenario 2 sebagai konfigurasi terbaik dalam penelitian ini tidak didasarkan pada signifikansi statistik semata, tetapi juga mempertimbangkan aspek praktis, yaitu kestabilan performa dan kemampuan model dalam mengenali kelas minoritas secara lebih seimbang.

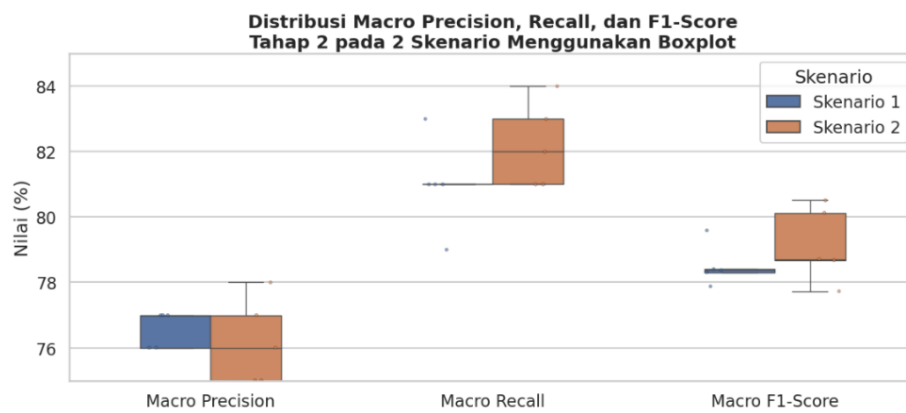
Oleh karena itu, dalam menentukan model terbaik, tidak cukup hanya mempertimbangkan nilai metrik evaluasi, tetapi juga perlu memperhatikan aspek lain seperti stabilitas performa, konsistensi hasil antar percobaan, serta kompleksitas model yang digunakan.

Visualisasi *boxplot* pada Gambar 4.7 menunjukkan distribusi nilai evaluasi model pada tiga skenario yang telah diuji, sehingga perbedaan performa antar skenario dapat terlihat dengan lebih jelas. Melalui grafik ini, pola perbedaan hasil pengujian juga dapat diamati untuk menilai ada atau tidaknya indikasi perbedaan

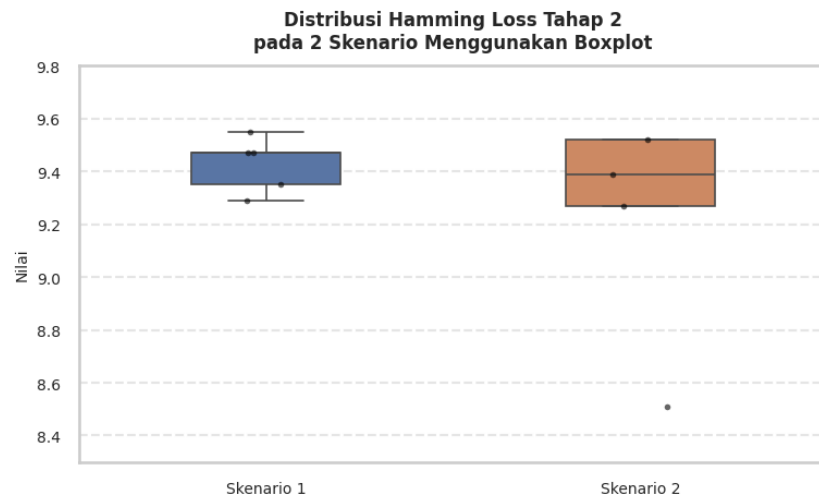
yang signifikan. Selain itu, *boxplot* membantu memperlihatkan tingkat kestabilan, konsistensi, serta keragaman hasil pengujian pada setiap metrik evaluasi, yang tidak dapat dijelaskan secara menyeluruh hanya berdasarkan nilai rata-rata dan standar deviasi.



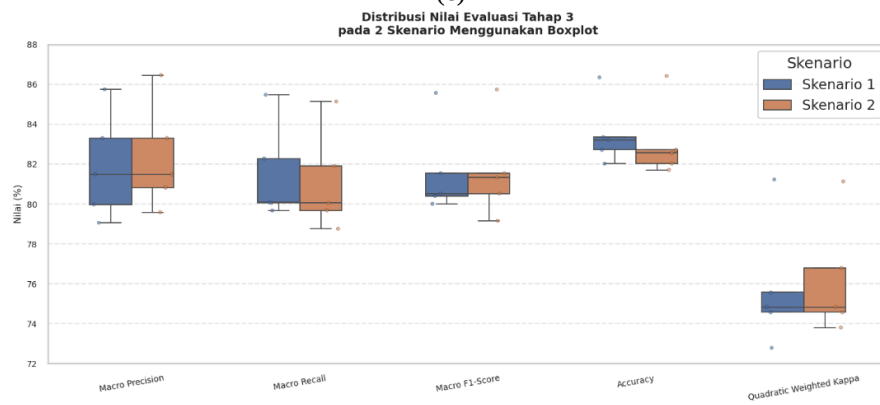
(a)



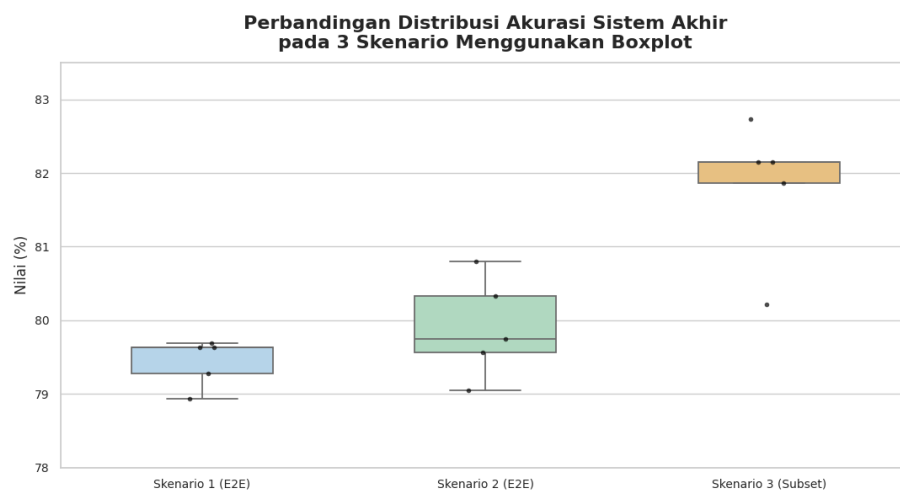
(b)



(c)



(d)



(e)

Gambar 4.7 *Boxplot* perbandingan skenario 1, 2, dan 3 (a) Tahap 1, (b) Tahap 2, (c) Tahap 2 *Hamming Loss*, (d) Tahap 3, (e) Akurasi sistem akhir antara 3 skenario

Berdasarkan Gambar 4.7 yang menunjukkan *boxplot* pada tahap klasifikasi biner, dapat dilihat bahwa performa model pada kedua skenario berada pada tingkat yang tinggi dan relatif stabil. Nilai *precision*, *F1-Score*, dan *accuracy* berada di atas 90% dengan sebaran yang sempit, yang menunjukkan bahwa model IndoBERTweet mampu membedakan antara ujaran kebencian dan non-ujaran kebencian dengan sangat baik. Namun demikian, nilai *recall* menunjukkan variasi yang lebih besar dibandingkan metrik lainnya, terutama pada Skenario 1 yang memiliki beberapa nilai lebih rendah (*outlier*). Hal ini mengindikasikan bahwa masih terdapat kasus *false negative*, di mana ujaran kebencian tidak berhasil terdeteksi oleh model, yang berpotensi mempengaruhi performa pada tahap selanjutnya dalam sistem klasifikasi hierarkis.

Selanjutnya, pada tahap klasifikasi multi-label, terlihat bahwa performa model mengalami penurunan dibandingkan tahap sebelumnya seiring dengan meningkatnya kompleksitas tugas. Nilai *Macro Precision* dan *Macro F1-Score* berada pada kisaran 76–80%, yang menunjukkan bahwa model cukup mampu mengidentifikasi lebih dari satu kategori ujaran kebencian dalam satu data. Nilai *Macro Recall* pada Skenario 2 tampak sedikit lebih tinggi dibandingkan Skenario 1, yang mengindikasikan adanya peningkatan kemampuan model dalam menangkap label yang relevan. Sementara itu, nilai *Hamming Loss* yang berada pada kisaran sekitar 9–10% (dalam penelitian ini disajikan dalam bentuk persentase) menunjukkan bahwa proporsi kesalahan prediksi label masih relatif rendah. Meskipun demikian, peningkatan performa pada Skenario 2 dibandingkan Skenario 1 relatif kecil, sehingga dapat dikatakan tidak memberikan pengaruh yang

signifikan. Hal ini menunjukkan bahwa penerapan data *augmentation* menggunakan *Masked Language Modeling* (MLM) belum mampu meningkatkan performa model secara optimal. Selain itu, sebaran nilai yang masih cukup lebar menunjukkan bahwa model masih mengalami kesulitan dalam menangani ketidakseimbangan distribusi label, terutama pada kategori yang jarang muncul.

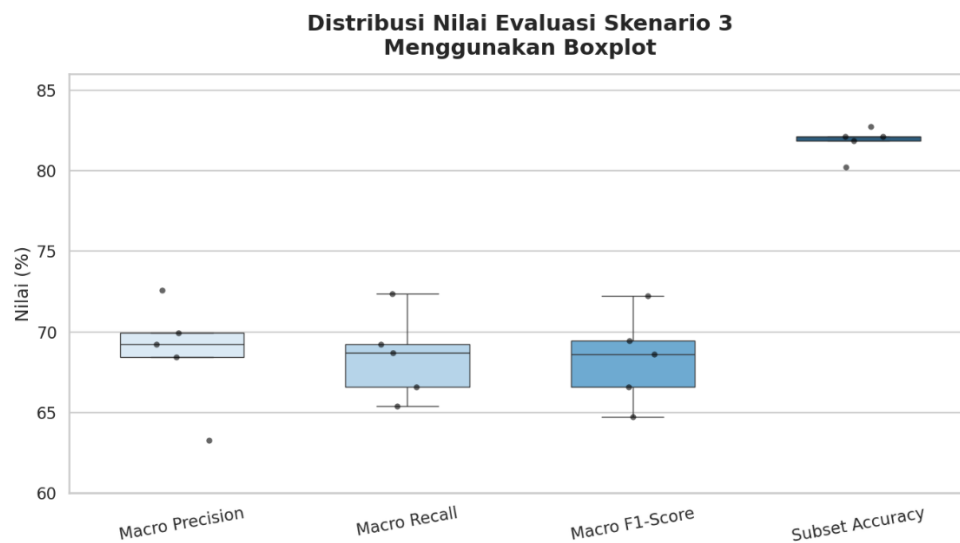
Pada *boxplot* tahap klasifikasi tingkat keparahan, terlihat bahwa performa model kembali menunjukkan kestabilan yang cukup baik. Nilai *Macro F1-Score* berada pada kisaran 80–82%, sedangkan *accuracy* mencapai sekitar 82–83%. Selain itu, nilai *Quadratic Weighted Kappa* (QWK) berada pada kisaran 74–76%, yang menunjukkan bahwa model mampu mempertahankan hubungan ordinal antar kelas keparahan dengan cukup baik. Hal ini menunjukkan bahwa pendekatan regresi ordinal (CORN) efektif dalam menangkap urutan tingkat keparahan ujaran kebencian. Meskipun demikian, perbedaan performa antara Skenario 1 dan Skenario 2 juga tidak menunjukkan peningkatan yang signifikan, yang kembali mengindikasikan bahwa pengaruh data *augmentation* terhadap tahap ini relatif terbatas.

Perbandingan akurasi sistem secara *end-to-end*, terlihat bahwa Skenario 1 menghasilkan akurasi pada kisaran 79–80%, sementara Skenario 2 hanya menunjukkan sedikit peningkatan menjadi sekitar 79–81%. Adapun Skenario 3 (subset) menghasilkan performa terbaik dengan akurasi mencapai kisaran 80–83%. Peningkatan yang relatif kecil antara Skenario 1 dan Skenario 2 menunjukkan bahwa penerapan teknik data *augmentation* menggunakan MLM tidak memberikan pengaruh yang signifikan terhadap peningkatan performa sistem secara

keseluruhan. Sebaliknya, peningkatan yang lebih jelas pada Skenario 3 menunjukkan bahwa kualitas dan distribusi data memiliki pengaruh yang lebih dominan dibandingkan teknik *augmentation* yang digunakan dalam penelitian ini.

Secara keseluruhan, hasil visualisasi *boxplot* pada setiap tahap menunjukkan bahwa model memiliki performa terbaik pada tahap klasifikasi biner, mengalami penurunan pada tahap multi-label, dan kembali stabil pada tahap klasifikasi tingkat keparahan. Hal ini menunjukkan bahwa tahap klasifikasi multi-label merupakan bagian yang paling menantang dan menjadi *bottleneck* utama dalam sistem, yang terutama disebabkan oleh ketidakseimbangan data serta kompleksitas dalam memahami konteks bahasa informal, implisit, dan sarkastik pada data media sosial. Selain itu, hasil penelitian juga menunjukkan bahwa peningkatan jumlah data melalui teknik augmentasi tidak selalu berbanding lurus dengan peningkatan performa model, sehingga diperlukan pendekatan lain yang lebih efektif dalam menangani permasalahan distribusi data dan kompleksitas bahasa.

Sedangkan pada Skenario 3 dengan pendekatan *flat*, model menunjukkan performa yang lebih tinggi, khususnya pada metrik accuracy, seperti yang ditunjukkan pada Gambar 4.8. Meskipun demikian, peningkatan ini tidak secara langsung mencerminkan kemampuan model dalam mendeteksi ujaran kebencian secara optimal.



Gambar 4.8 Boxplot Pada Skenario 3

Berdasarkan Gambar 4.8, pendekatan *flat* pada Skenario 3 memiliki nilai *accuracy* yang lebih tinggi dibandingkan dua skenario sebelumnya. Hal ini terjadi karena pada pendekatan *flat*, seluruh prediksi termasuk kelas *non-hate speech* (NON-HS) turut diperhitungkan dalam evaluasi, sehingga jumlah prediksi benar menjadi lebih besar dan berdampak pada peningkatan nilai *accuracy*. Berbeda dengan pendekatan hierarkis yang lebih berfokus pada data ujaran kebencian (HS), sehingga perhitungan performanya menjadi lebih ketat terhadap kasus yang relevan.

Namun demikian, nilai *recall* pada Skenario 3 mengalami penurunan, yang menunjukkan bahwa model cenderung gagal dalam mendeteksi sebagian data ujaran kebencian yang sebenarnya ada (*false negative* meningkat). Penurunan *recall* ini mengindikasikan bahwa meskipun model terlihat memiliki akurasi yang tinggi, kemampuan model dalam menangkap seluruh kasus ujaran kebencian menjadi berkurang. Dengan kata lain, model cenderung lebih “aman” dalam memprediksi NON-HS, tetapi mengorbankan sensitivitas terhadap HS.

Kondisi ini berdampak pada ketidakseimbangan performa model, di mana nilai *accuracy* yang tinggi tidak sepenuhnya mencerminkan kemampuan model dalam mendeteksi ujaran kebencian secara menyeluruh. Oleh karena itu, pendekatan *flat* pada Skenario 3 meskipun unggul dari sisi *accuracy*, kurang optimal jika digunakan sebagai sistem deteksi ujaran kebencian yang membutuhkan kemampuan identifikasi yang sensitif terhadap kelas HS.

Secara keseluruhan, keunggulan pendekatan klasifikasi hierarkis dalam penelitian ini ditunjukkan pada beberapa aspek. Pertama, pendekatan klasifikasi hierarkis mampu membagi proses klasifikasi menjadi beberapa tahap yang lebih spesifik. Pada Tahap 1, model hanya berfokus pada klasifikasi biner antara ujaran kebencian dan non-ujaran kebencian, sehingga kompleksitas tugas menjadi lebih sederhana dibandingkan klasifikasi seluruh label secara langsung. Hal ini terlihat dari performa Tahap 1 yang memperoleh nilai *F1-Score* dan *Accuracy* sebesar 90,37% dan 90,67%, yang menunjukkan bahwa model mampu mengenali pola ujaran kebencian dengan baik pada tahap awal klasifikasi.

Kedua, pendekatan hierarkis memudahkan proses analisis kesalahan karena performa model dapat diamati pada setiap tahap secara terpisah. Contohnya, hasil evaluasi menunjukkan bahwa Tahap 2 memiliki performa yang lebih rendah dibandingkan tahap lainnya secara deskriptif, yang ditunjukkan oleh nilai *Macro Precision* sebesar 76%, *Macro Recall* sebesar 81–82%, serta *Macro F1-Score* sebesar 78,41% pada Skenario 1 dan 78,68% pada Skenario 2. Nilai tersebut lebih rendah dibandingkan Tahap 1 yang memiliki *F1-Score* sebesar 90,67% selanjutnya Tahap 3 yang memiliki *Macro F1-Score* sebesar 80,40% pada Skenario 1 dan

81,34% pada Skenario 2, sehingga mengindikasikan bahwa klasifikasi multi-label menjadi bottleneck utama dalam sistem. Kondisi ini dipengaruhi oleh ketidakseimbangan distribusi data antar label serta tingkat kompleksitas tugas yang lebih tinggi dibandingkan klasifikasi biner maupun ordinal. Selain itu, kesalahan pada Tahap 1 juga berdampak pada tahap berikutnya karena data yang salah diklasifikasikan tidak dapat diteruskan ke proses selanjutnya.

Ketiga, dibandingkan pendekatan *flat classification* pada Skenario 3, pendekatan hierarkis memiliki proses klasifikasi yang lebih terstruktur karena prediksi dilakukan secara bertahap sesuai tugas pada masing-masing level klasifikasi. Pada pendekatan *flat classification*, model harus memprediksi seluruh label secara langsung dalam satu tahap, sedangkan pada pendekatan hierarkis proses klasifikasi dibagi ke dalam beberapa tahap, yaitu klasifikasi biner, klasifikasi multi-label, dan klasifikasi tingkat keparahan. Struktur tersebut membantu proses klasifikasi menjadi lebih sistematis serta memudahkan identifikasi sumber kesalahan pada setiap tahap pengujian. Dengan demikian, pendekatan hierarkis tidak hanya memberikan performa yang kompetitif, tetapi juga meningkatkan interpretabilitas dan pemahaman terhadap proses klasifikasi secara keseluruhan.

4.2.2 Analisis Pengaruh *Threshold*

Dalam penelitian ini, nilai *threshold* sebesar 0,5 digunakan sebagai batas keputusan dalam klasifikasi multi-label. Pemilihan nilai ini didasarkan pada praktik umum dalam penggunaan fungsi sigmoid, di mana probabilitas $\geq 0,5$ dikategorikan sebagai kelas positif. Berdasarkan hasil pengujian, penggunaan *threshold* ini sudah

mampu memberikan keseimbangan yang cukup baik antara *precision* dan *recall*, sehingga dapat dianggap memadai untuk penelitian ini.

Meskipun demikian, masih ditemukan beberapa kasus *false negative* pada label tertentu, yang menunjukkan bahwa nilai *threshold* 0,5 belum sepenuhnya optimal untuk seluruh label. Hal ini dipengaruhi oleh perbedaan distribusi probabilitas antar label, terutama pada label dengan jumlah data yang lebih sedikit, yang cenderung menghasilkan nilai probabilitas lebih rendah. Akibatnya, beberapa label yang seharusnya terdeteksi tidak memenuhi batas *threshold* yang ditentukan.

Oleh karena itu, pada penelitian selanjutnya dapat dipertimbangkan penggunaan *threshold* yang berbeda untuk setiap label (*label-specific threshold*) atau melakukan optimasi *threshold* berdasarkan metrik evaluasi tertentu seperti *F1-Score*, guna memperoleh keseimbangan yang lebih optimal antara *precision* dan *recall*.

4.2.3 Analisis Pengaruh *Balancing Data* Pada Model

Penggunaan teknik *balancing data* melalui augmentasi berbasis *Masked Language Modeling* (MLM) pada Skenario 2 bertujuan untuk mengatasi ketidakseimbangan distribusi label, khususnya pada tahap klasifikasi multi-label. Secara deskriptif, penerapan *balancing data* menunjukkan adanya peningkatan performa dibandingkan Skenario 1. Hal ini terlihat dari kenaikan nilai *macro F1-Score* sebesar 2,20% serta penurunan *Hamming Loss* sebesar 0,57%. Peningkatan ini mengindikasikan bahwa augmentasi data membantu model dalam mengenali label minoritas dengan lebih baik.

Namun demikian, berdasarkan hasil uji statistik menggunakan *paired t-test*, perbedaan performa antara Skenario 1 dan Skenario 2 tidak menunjukkan signifikansi secara statistik ($p\text{-value} > 0,05$). Hal ini menunjukkan bahwa peningkatan yang terjadi belum cukup kuat untuk dianggap sebagai peningkatan yang konsisten secara statistik, melainkan masih dapat dipengaruhi oleh variasi antar percobaan.

Pada tahap klasifikasi multi-label, *balancing* data tetap memberikan dampak praktis dengan membantu distribusi label menjadi lebih merata. Meskipun demikian, tahap ini masih menjadi bottleneck dalam sistem karena kompleksitas tugas multi-label dan tingginya ketidakseimbangan data. Dengan kata lain, *balancing* data membantu, tetapi belum sepenuhnya menyelesaikan permasalahan tersebut.

Sementara itu, pada tahap klasifikasi tingkat keparahan, pengaruh *balancing* data tidak terlihat secara signifikan terhadap performa model. Hal ini disebabkan karena pendekatan ordinal menggunakan metode CORN lebih berfokus pada hubungan antar level keparahan dibandingkan distribusi jumlah data pada masing-masing kelas. Akibatnya, peningkatan dari *balancing* data pada tahap sebelumnya tidak secara langsung berdampak besar pada tahap ini.

Secara keseluruhan, *balancing* data melalui MLM memberikan kontribusi positif secara empiris terhadap performa model, khususnya dalam meningkatkan kemampuan deteksi label minoritas. Namun, berdasarkan analisis statistik, pengaruh tersebut belum signifikan. Oleh karena itu, *balancing* data dapat dianggap sebagai teknik pendukung yang membantu peningkatan performa, tetapi masih

perlu dikombinasikan dengan strategi lain untuk menghasilkan peningkatan yang lebih konsisten dan signifikan.

4.2.4 Analisis Kesalahan (*Error Analysis*)

Analisis kesalahan dilakukan untuk mengidentifikasi sumber utama kesalahan model pada skenario terbaik (Skenario 2), serta memahami bagaimana kesalahan tersebut mempengaruhi performa sistem secara keseluruhan dalam pendekatan klasifikasi hierarkis. Berikut contoh error pada evaluasi E2E pada Gambar 4.9.

	Teks	T1_GT	T1_Pred	T2_GT	T2_Pred	T3_GT	T3_Pred	Status
0	ga caya sama mulut buaya mu bwang	1	0	Individual, Other	-	0	NaN	Error_T1
1	NaN	1	0	Group, Other	-	1	NaN	Error_T1
2	buat film muslim pilih ahok mati nggak di shol...	1	1	Individual, Other	Religion, Individual	0	0.0	Error_T2
3	nantang buka dadalah eh nyebut setanlah emang ...	1	1	Group, Physical	Individual, Other	1	0.0	Error_T2
4	dpr mpr segera ambil tindakan pecat dengan tid...	1	1	Individual, Other	Individual, Other	0	2.0	Error_T3
5	kalah provokasi lengeserkan jokowigoe ikut jug...	1	1	Individual, Other	Individual, Other	0	2.0	Error_T3

Gambar 4.9 Contoh *Error Analysis* secara keseluruhan

Analisis kesalahan pada gambar 4.9 dilakukan untuk mengidentifikasi sumber utama kesalahan model pada skenario terbaik (Skenario 2), serta memahami bagaimana kesalahan tersebut mempengaruhi performa sistem secara keseluruhan dalam pendekatan klasifikasi hierarkis. Berdasarkan hasil evaluasi *end-to-end*, ditemukan bahwa kesalahan terjadi pada tahap klasifikasi biner (Tahap 1), khususnya dalam bentuk *false negative*. Pada kasus ini, data yang seharusnya termasuk ujaran kebencian justru diklasifikasikan sebagai non-ujaran kebencian, sehingga tidak dapat dilanjutkan ke tahap berikutnya. Sebagai contoh, pada teks “ga caya sama mulut buaya mu bwang”, model gagal mengidentifikasi adanya ujaran kebencian meskipun secara konteks mengandung unsur serangan terhadap individu.

Kesalahan ini menunjukkan bahwa model masih mengalami keterbatasan dalam memahami konteks implisit dan bahasa informal, seperti penggunaan kata kiasan atau ungkapan tidak langsung. Berbeda dengan ujaran kebencian yang bersifat eksplisit, kasus seperti ini membutuhkan pemahaman semantik yang lebih dalam, yang belum sepenuhnya mampu ditangkap oleh model.

Selain itu, kesalahan pada Tahap 1 memiliki dampak yang signifikan terhadap keseluruhan sistem karena menyebabkan data tidak diproses lebih lanjut pada Tahap 2 dan Tahap 3. Hal ini mengindikasikan adanya efek propagasi kesalahan dalam pendekatan hierarkis, di mana kesalahan pada tahap awal akan mengakibatkan hilangnya seluruh informasi pada tahap berikutnya. Dengan kata lain, satu kesalahan pada tahap awal dapat menghasilkan kesalahan total pada output akhir sistem.

Pada tahap klasifikasi multi-label (Tahap 2), mekanisme evaluasi berbeda dengan evaluasi *end-to-end*. Pada evaluasi *end-to-end*, suatu prediksi dianggap benar hanya jika seluruh tahapan (klasifikasi biner, kategori, dan tingkat keparahan) menghasilkan output yang sesuai dengan *ground truth*. Dengan kata lain, pendekatan ini bersifat ketat (*strict*), di mana kesalahan pada satu bagian saja akan menyebabkan keseluruhan prediksi dianggap salah.

Sebaliknya, pada tahap klasifikasi multi-label, evaluasi tidak menggunakan pendekatan subset secara langsung, melainkan mempertimbangkan setiap label secara independen. Hal ini berarti bahwa model masih dapat memberikan prediksi yang sebagian benar meskipun tidak seluruh label berhasil dideteksi. Sebagai contoh, jika suatu data memiliki dua label kategori (*Individual* dan *Other*), dan

model hanya berhasil memprediksi satu di antaranya, maka prediksi tersebut tetap memiliki nilai kebenaran parsial. Karakteristik ini menyebabkan kesalahan pada tahap multi-label tidak selalu bersifat absolut, melainkan dapat berupa:

1. *False Negative* (FN): label yang seharusnya ada tetapi tidak terdeteksi oleh model
2. *False Positive* (FP): label yang diprediksi oleh model tetapi tidak terdapat pada *ground truth*
3. Parsial benar: sebagian label berhasil diprediksi dengan tepat

Berikut contoh error prediksi pada tahap 2 multi-label pada Gambar 4.10.

index	Teks	GT_Label	Pred_Label	Benar	Miss (FN)	Salah Tambah (FP)
1	selamat hari kartini	-	Individual, Other	-	-	Individual, Other
6	korsa pernyataan moeldoko bukti istana takut dengan kebangkitan umat islam	-	Individual, Religion	-	-	Individual, Religion
9	buat film muslim pilih ahok mati nggak di sholatin dijamin laris	Individual, Other	Individual, Religion	Individual	Other	Religion
13	sama aja lh laemeskipun sdh turun menurun tp ideologi pki silu melekat di jiwa mreka	-	Race, Group	-	-	Race, Group
16	yg cinta orba pasti meng idola kan prabowo amp hary tanoe d	Individual, Other	Group, Other	Other	Individual	Group
17	nantang buka dadalah eh nyebut setanlah emang sudah pada habis kamus ngelawan kenegarawanan pak joko widodo presiden ni ke pada makin stres nie hati kena stoke atau serangan jantung nanti rakyat indonesia kehilangan badut politik merdeka	Group, Physical	Individual, Other	-	Group, Physical	Individual, Other
18	jongos penista agama ahok sok sok ngong saramajikan nya penebar sara astagfirillah tobat	Individual, Other	Individual, Religion	Individual	Other	Religion
20	gak vital gimanaudah dari kapan inter gak punya bek kanan yg bisa jd pembeda terakhir itu ya era maicon dambrosio juga baru setelah kedatangan cancelo bagusnyanhmberarti kesimpulannya yg geblek ausilio ini xfxxf	Individual, Other	Group, Other	Other	Individual	Group
21	nahklo pki alias partai korupsi indonesia perlu diganyang juga ga	Group, Other	Race, Group	Group	Other	Race

Gambar 4.10 Contoh *Error Analysis* Pada Tahap 2 Multi-label

Pada Gambar 4.10 berdasarkan hasil analisis, ditemukan bahwa sebagian besar kesalahan pada tahap ini berupa *false negative*, di mana model tidak mampu mendeteksi seluruh label yang relevan dalam satu data. Hal ini menunjukkan bahwa model cenderung belum optimal dalam menangkap kompleksitas hubungan antar kategori dalam satu teks, terutama pada data dengan lebih dari satu label.

Selain itu, terdapat juga kasus *false positive* yang menunjukkan bahwa model cenderung melakukan *over prediction* terhadap kategori tertentu. Hal ini mengindikasikan bahwa model masih mengalami kesulitan dalam membedakan konteks antar kategori yang memiliki kemiripan semantik.

Perlu diperhatikan bahwa meskipun terjadi kesalahan pada tahap multi-label, data tetap dapat diproses ke tahap berikutnya (klasifikasi tingkat keparahan). Hal ini berbeda dengan tahap klasifikasi biner, di mana kesalahan akan menghentikan seluruh proses. Dengan demikian, kesalahan pada tahap multi-label memiliki dampak yang lebih terbatas dibandingkan kesalahan pada tahap awal, namun tetap berpengaruh terhadap kualitas prediksi akhir.

Secara keseluruhan, hasil analisis menunjukkan bahwa tahap multi-label merupakan bagian yang kompleks dalam sistem, karena melibatkan kemungkinan lebih dari satu label dalam satu data. Oleh karena itu, evaluasi pada tahap ini lebih tepat menggunakan metrik seperti *Macro F1* dan *Hamming Loss*, yang mampu merepresentasikan performa model secara lebih adil dibandingkan akurasi berbasis subset.

Pada tahap klasifikasi tingkat keparahan (Tahap 3), kesalahan relatif lebih kecil dibandingkan tahap sebelumnya. Namun demikian, model masih mengalami kesulitan dalam membedakan kelas dengan tingkat keparahan yang berdekatan, khususnya antara kelas sedang dan kuat. Hal ini disebabkan oleh sifat data yang ambigu serta batas antar kelas yang tidak selalu jelas secara linguistik.

Secara keseluruhan, hasil analisis menunjukkan bahwa sumber utama kesalahan dalam penelitian ini berasal dari tahap klasifikasi multi-label yang memicu propagasi kesalahan dalam sistem hierarkis. Faktor lain yang turut mempengaruhi adalah penggunaan bahasa informal, ambiguitas konteks, serta ketidakseimbangan distribusi data. Oleh karena itu, peningkatan performa model di masa mendatang perlu difokuskan pada peningkatan pemahaman konteks bahasa

informal, optimalisasi klasifikasi multi-label, serta perbaikan mekanisme deteksi awal untuk meminimalkan dampak kesalahan berantai dalam sistem.

Penelitian ini bertujuan untuk menjaga etika dalam berucap di media sosial sebagai upaya mencegah terjadinya dampak sosial yang bersifat destruktif akibat ujaran kebencian. Melalui pengembangan sistem deteksi ujaran kebencian, penelitian ini diharapkan dapat berkontribusi dalam menciptakan ruang digital yang aman dan kondusif, sehingga pengguna media sosial, baik individu maupun pemilik akun atau platform, dapat mengekspresikan pendapatnya secara bebas dan bertanggung jawab tanpa menimbulkan persepsi negatif, penghinaan, atau konflik sosial.

Dalam konteks ini, Islam juga menekankan pentingnya ketelitian dalam menerima dan menilai suatu informasi, sebagaimana dijelaskan dalam Al-Qur'an Surah Al-Hujurat ayat 6 yang memerintahkan untuk melakukan verifikasi sebelum mengambil keputusan.

يَا أَيُّهَا الَّذِينَ آمَنُوا إِنْ جَاءَكُمْ فَاسِقٌ بِنَبَأٍ فَتَبَيَّنُوا أَنْ تُصِيبُوا قَوْمًا بِجَهَالَةٍ فَتُصْحَبُوا عَلَىٰ مَا فَعَلْتُمْ نَادِمِينَ

“Wahai orang-orang yang beriman! Jika seseorang yang fasik datang kepadamu membawa suatu berita, maka telitilah kebenarannya, agar kamu tidak mencelakakan suatu kaum karena kebodohan (kecerobohan), yang akhirnya kamu menyesali perbuatanmu itu.” (QS Al-Hujurat:6)

Menurut tafsir Ibnu Katsir, ayat ini menjelaskan bahwa setiap informasi yang datang, terutama dari sumber yang tidak jelas, harus diteliti terlebih dahulu agar tidak menimbulkan kesalahan dalam pengambilan keputusan yang dapat merugikan pihak lain. Prinsip ini sejalan dengan proses evaluasi dalam penelitian ini, di mana

hasil prediksi model tidak langsung diterima, melainkan diuji menggunakan berbagai metrik untuk memastikan keakuratan dan keandalannya

Prinsip ini sejalan dengan proses evaluasi dalam penelitian ini, di mana hasil prediksi model tidak langsung diterima, melainkan diuji menggunakan berbagai metrik untuk memastikan keakuratan dan keandalannya.

Ujaran kebencian sering kali berasal dari tauhid atau keimanan yang tidak utuh. Jika seseorang sungguh-sungguh beriman kepada Allah, ia tidak akan menyebarkan kebencian, karena perbuatan tersebut bertentangan dengan ajaran-Nya. Prinsip menjaga etika komunikasi ini sejalan dengan ajaran Islam yang menegaskan larangan saling merendahkan, mencela, dan memberi julukan buruk kepada sesama, sebagaimana dijelaskan dalam Al-Qur'an surah Al-Hujurat ayat 11 berikut:

يَا أَيُّهَا الَّذِينَ آمَنُوا لَا يَسْخَرُ قَوْمٌ مِّنْ قَوْمٍ عَسَىٰ أَن يَكُونَ خَيْرًا مِّنْهُمْ وَلَا نِسَاءٌ مِّنْ نِّسَاءٍ عَسَىٰ أَن يَكُنَّ خَيْرًا مِّنْهُنَّ وَلَا تَلْمِزُوا أَنْفُسَكُمْ وَلَا تَنَابَزُوا بِالْأَلْقَابِ بِئْسَ الْإِسْمُ الْفُسُوقُ بَعْدَ الْإِيمَانِ وَمَنْ لَّمْ يَتُبْ فَأُولَٰئِكَ هُمُ الظَّالِمُونَ

“Wahai orang-orang yang beriman, janganlah suatu kaum mengolok-olok kaum yang lain (karena) boleh jadi mereka (yang diolok-olokkan itu) lebih baik daripada mereka (yang mengolok-olok) dan jangan pula perempuan-perempuan (mengolok-olok) perempuan lain (karena) boleh jadi perempuan (yang diolok-olok itu) lebih baik daripada perempuan (yang mengolok-olok). Janganlah kamu saling mencela dan saling memanggil dengan julukan yang buruk. Seburuk-buruk panggilan adalah (panggilan) fasik setelah beriman. Siapa yang tidak bertobat, mereka itulah orang-orang zalim.”(Q.S. Al-Hujurat ayat 11)

Menurut Tafsir Ibnu Katsir, ayat tersebut mengajarkan untuk saling menghormati serta larangan menghina dan merendahkan orang lain, karena hal tersebut sudah jelas haram. Terkadang karena orang yang diremehkan justru mulia disisi Allah SWT. Ibnu Katsir juga menjelaskan bahwa mencela orang dan menghina mereka dengan sewenang-wenang dan berjalan kesana kemari untuk mengadu domba itu termasuk celaan dalam bentuk ucapan. Oleh karena itu Allah

SWT berfirman وَلَا تَلْمِزُوا أَنْفُسَكُمْ “dan janganlah kamu mencela dirimu sendiri” karena pada hakikatnya mencela orang lain sama dengan mencela diri sendiri. Allah berfirman pada وَلَا تَقْتُلُوا أَنْفُسَكُمْ “dan janganlah kamu membunuh dirimu sendiri” (Surah An-Nisa: 29).

Menjaga etika dalam berucap juga ditegaskan dalam hadist Nabi Muhammad SAW yang diriwayatkan oleh Imam Bukhori dan Muslim :

مَنْ كَانَ يُؤْمِنُ بِاللَّهِ وَالْيَوْمِ الْآخِرِ فَلْيُكُنْ خَيْرًا أَوْ لِيَصْنَعْ

“Barang siapa beriman kepada Allah dan hari akhir, maka hendaklah ia berkata baik atau diam.” (HR. Bukhari dan Muslim)

Hadis ini menegaskan bahwa menjaga lisan merupakan wujud keimanan. Dalam konteks media sosial, penerapan teknologi deteksi ujaran kebencian dapat dipandang sebagai ikhtiar manusia untuk mematuhi perintah Allah SWT dalam menjaga ucapan dan menghindari perbuatan yang dilarang. Dengan demikian, penelitian ini mendukung upaya pengamalan nilai keislaman dalam ruang digital sebagai bentuk tanggung jawab kepada perintah Allah SWT.

Ujaran kebencian di media sosial memiliki dampak langsung terhadap hubungan sosial, seperti memicu konflik, diskriminasi, perpecahan, serta merusak keharmonisan antarindividu maupun kelompok. Oleh karena itu, menjaga etika komunikasi merupakan bagian dari kewajiban manusia dalam membina hubungan yang baik dengan sesama manusia (*habluminannas*). Prinsip menjaga hubungan sesama manusia (*habluminannas*) juga ditegaskan dalam hadis Nabi Muhammad SAW yang menyatakan:

الْمُسْلِمُ مَنْ سَلِمَ الْمُسْلِمُونَ مِنْ لِسَانِهِ وَيَدِهِ

"Seorang muslim (yang sejati) adalah orang yang kaum muslimin selamat dari gangguan lisan dan tangannya." (HR. Bukhori dan Muslim)

Hadis tersebut menegaskan bahwa lisan memiliki peran penting dalam menjaga keharmonisan sosial. Gangguan yang ditimbulkan melalui ucapan, termasuk ujaran kebencian di media sosial, berpotensi melukai individu maupun kelompok serta merusak hubungan sosial. Penelitian ini berkontribusi dalam menjaga hubungan antarmanusia dengan menghadirkan sistem deteksi ujaran kebencian yang dapat mengidentifikasi konten bermuatan penghinaan, provokasi, dan kekerasan verbal. Dengan adanya sistem ini, diharapkan interaksi di ruang digital dapat berlangsung secara lebih aman, santun, dan bertanggung jawab. Upaya ini sejalan dengan prinsip *amar ma'ruf nahi munkar* dalam Islam, yaitu mencegah terjadinya kemudharatan sosial dan menjaga keharmonisan kehidupan sosial.

Ruang digital dan media sosial dapat dipandang sebagai bagian dari lingkungan sosial yang dibentuk dan digunakan oleh manusia. Penyebaran ujaran kebencian secara masif dapat menciptakan lingkungan digital yang tidak sehat, penuh konflik, dan merusak kenyamanan serta keamanan pengguna. Oleh karena itu, menjaga kualitas dan keseimbangan ruang digital merupakan bagian dari tanggung jawab manusia terhadap lingkungannya (*hablumminal 'alam*).

Islam melarang segala bentuk perbuatan yang menimbulkan kerusakan, sebagaimana firman Allah SWT dalam Al-Qur'an surah Al-A'raf ayat 56:

وَلَا تُفْسِدُوا فِي الْأَرْضِ بَعْدَ إِصْلَاحِهَا وَادْعُوهُ خَوْفًا وَطَمَعًا إِنَّ رَحْمَتَ اللَّهِ قَرِيبٌ مِنَ الْمُحْسِنِينَ

“Dan janganlah kamu membuat kerusakan di muka bumi setelah (Allah) memperbaikinya. Berdoalah kepada-Nya dengan rasa takut dan harap. Sesungguhnya rahmat Allah sangat dekat kepada orang-orang yang berbuat kebaikan.” (Q.S. Al-A’raf: 56)

Menurut Tafsir Ibnu Katsir Allah Swt melarang perbuatan yang menimbulkan kerusakan di muka bumi dan hal-hal yang membahayakan kelestariannya sesudah diperbaiki. Karena sesungguhnya apabila segala sesuatunya berjalan sesuai dengan kelestariannya, kemudian terjadilah pengrusakan padanya, hal tersebut akan membahayakan semua hamba Allah.

Ayat tersebut menegaskan larangan melakukan perbuatan yang merusak keseimbangan dan ketertiban dalam kehidupan. Kerusakan tidak hanya bersifat fisik, tetapi juga mencakup kerusakan sosial dan psikologis. Dalam konteks media sosial, ujaran kebencian dapat dipandang sebagai bentuk perbuatan yang mencemari ekosistem komunikasi dan merusak tatanan lingkungan digital.

Penelitian ini berkontribusi dalam menjaga lingkungan digital yang sehat melalui pengembangan sistem deteksi ujaran kebencian. Sistem ini berperan sebagai alat bantu untuk mengidentifikasi dan meminimalkan penyebaran konten bermuatan negatif, sehingga ruang digital dapat dimanfaatkan secara lebih positif, aman, dan berkelanjutan. Dengan demikian, pemanfaatan teknologi dalam penelitian ini sejalan dengan nilai-nilai Islam dalam menjaga keseimbangan dan kemaslahatan lingkungan sebagai amanah yang harus dijaga bersama.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa penerapan metode klasifikasi hierarkis multi-label menggunakan model IndoBERTweet mampu mengidentifikasi ujaran kebencian beserta kategori dan tingkat keparahannya secara bertingkat dengan performa yang cukup baik pada data berbahasa Indonesia.

Pada tahap klasifikasi biner (Tahap 1), model memperoleh *Accuracy* sebesar 90,67% dan *F1-Score* sebesar 90,37%. Pada tahap klasifikasi multi-label (Tahap 2), model mencapai *Macro F1-Score* sebesar 78,68% dan *Hamming Loss* sebesar 9,39%. Pada tahap klasifikasi ordinal tingkat keparahan (Tahap 3), metode CORN memperoleh *Accuracy* sebesar 82,56% dan *Quadratic Weighted Kappa* sebesar 76,80%. Secara keseluruhan, evaluasi *end-to-end system* klasifikasi hierarkis menghasilkan performa sebesar 80,33%, yang menunjukkan bahwa pendekatan hierarkis mampu mengidentifikasi ujaran kebencian, kategori ujaran kebencian, dan tingkat keparahannya dengan performa yang cukup baik.

Selain itu, hasil eksperimen menunjukkan bahwa penerapan *balancing* data mampu meningkatkan performa klasifikasi pada beberapa skenario pengujian. Sementara itu, augmentasi berbasis *Masked Language Modeling* (MLM) belum memberikan peningkatan performa yang signifikan secara statistik berdasarkan hasil uji *paired t-test*.

Hasil tersebut menunjukkan bahwa pendekatan hierarkis dapat digunakan sebagai alternatif yang efektif untuk memodelkan proses deteksi ujaran kebencian secara bertahap, mulai dari identifikasi ujaran kebencian, penentuan kategori, hingga prediksi tingkat keparahan. Dengan demikian, metode klasifikasi hierarkis multi-label berbasis IndoBERTweet memiliki potensi untuk mendukung pengembangan sistem deteksi ujaran kebencian yang lebih terstruktur pada data media sosial berbahasa Indonesia.

5.2 Saran

Berdasarkan hasil penelitian yang telah dilakukan, beberapa saran untuk penelitian selanjutnya adalah sebagai berikut:

1. Perbaiki ketidakseimbangan data pada tahap multi-label.

Tahap multi-label masih menjadi *bottleneck* akibat distribusi data yang tidak seimbang. Oleh karena itu, dapat digunakan pendekatan seperti *class weighting* atau *focal loss* agar model lebih mampu mengenali label minoritas.

2. Penerapan *iterative stratification* pada tahap multi-label.

Penelitian selanjutnya dapat menerapkan *iterative stratification* pada tahap multi-label untuk menjaga distribusi label pada data pelatihan dan pengujian serta meningkatkan representativitas hasil evaluasi.

3. Pengurangan propagasi kesalahan pada pendekatan hierarkis.

Untuk mengurangi ketergantungan antar tahap, dapat digunakan pendekatan *multi-task learning* atau model *end-to-end* agar proses klasifikasi dipelajari secara simultan dan lebih *robust*.

4. Pengembangan teknik augmentasi yang lebih efektif.

Mengingat augmentasi berbasis MLM belum memberikan peningkatan yang signifikan secara statistik ($p\text{-value} > 0,05$), penelitian selanjutnya disarankan untuk mengombinasikan metode augmentasi lain yang lebih variatif. Hal ini bertujuan agar peningkatan performa tidak hanya bersifat deskriptif, tetapi juga signifikan secara statistik.

5. Pengembangan variasi skenario dan desain eksperimen.

Penelitian ini masih memiliki variasi skenario yang terbatas karena perbedaan antar skenario hanya berfokus pada balancing data. Penelitian selanjutnya disarankan mengeksplorasi variasi arsitektur model, *hyperparameter*, maupun strategi pelatihan agar pengaruh setiap pendekatan dapat dianalisis secara lebih mendalam.

DAFTAR PUSTAKA

- Arifin, M., & Mahdiana, D. (2024). IMPLEMENTATION OF DEEP LEARNING MODELS IN HATE SPEECH DETECTION ON TWITTER USING AN NATURAL LANGUAGE PROCESSING APPROACH. *Jurnal Teknik Informatika (Jutif)*, 5(5), 1257–1266. <https://doi.org/10.52436/1.jutif.2024.5.5.2043>
- Ayele, A. A., Jalew, E. A., Ali, A. C., Yimam, S. M., & Biemann, C. (2024). *Exploring Boundaries and Intensities in Offensive and Hate Speech: Unveiling the Complex Spectrum of Social Media Discourse*.
- Bagestra, R., Misbullah, A., Zulfan, Z., Rasudin, R., Farsiah, L., & Nazhifah, S. A. (2024). Performance Assessment of Machine Learning and Transformer Models for Indonesian Multi-Label Hate Speech Detection. *Infolitika Journal of Data Science*, 2(2), 62–71. <https://doi.org/10.60084/ijds.v2i2.235>
- Febryanti, D. R., Mahmud, Z. K., Putri, S. V., Ovalia, F. M., & Sekarwangi, Y. (2022). Tipologi Hate Speech di Twitter Terkait Kebijakan Pemerintah Selama Pandemi COVID-19. *Jurnal Komunikasi Global*, 11(2), 274–299. <https://doi.org/10.24815/jkg.v11i2.26733>
- G. Ambar Wulan, R. G. M. P. M. (2021). Pencegahan Kejahatan Ujaran Kebencian di Indonesia. *Jurnal Ilmu Kepolisian*, 14(3), 19. <https://doi.org/10.35879/jik.v14i3.278>
- Guo, Y., Liu, Y., Bakker, E. M., Guo, Y., & Lew, M. S. (2018). CNN-RNN: A large-scale hierarchical image classification framework. *Multimedia Tools and Applications*, 77(8), 10251–10271. <https://doi.org/10.1007/s11042-017-5443-x>
- Haerudin, H., Suryadi, I., & Helmawati, H. (2025). Ethics of Communication on Social Media in The Perspective of The Interpretation of Ibn Kathir. *Studia Quranika*, 10(1), 113–130. <https://doi.org/10.21111/studiaquranika.v10i1.14805>
- Hendrawan, R., Adiwijaya, & Al Faraby, S. (2020). Multilabel Classification of Hate Speech and Abusive Words on Indonesian Twitter Social Media. *2020 International Conference on Data Science and Its Applications (ICoDSA)*, 1–7. <https://doi.org/10.1109/ICoDSA50139.2020.9212962>
- Henning, S., Beluch, W., Fraser, A., & Friedrich, A. (2023). A Survey of Methods for Addressing Class Imbalance in Deep-Learning Based Natural Language Processing. *Proceedings of the 17th Conference of the European Chapter*

of the Association for Computational Linguistics, 523–540.
<https://doi.org/10.18653/v1/2023.eacl-main.38>

Ibrohim, M. O., & Budi, I. (2019). Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter. *Proceedings of the Third Workshop on Abusive Language Online*, 46–57. <https://doi.org/10.18653/v1/W19-3506>

Ibrohim, M. O., & Budi, I. (2023). Hate speech and abusive language detection in Indonesian social media: Progress and challenges. *Heliyon*, 9(8), e18647. <https://doi.org/10.1016/j.heliyon.2023.e18647>

Koto, F., Lau, J. H., & Baldwin, T. (2021). IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10660–10668. <https://doi.org/10.18653/v1/2021.emnlp-main.833>

Kusuma, J. F., & Chowanda, A. (2023). Indonesian Hate Speech Detection Using IndoBERTweet and BiLSTM on Twitter. *JOIV : International Journal on Informatics Visualization*, 7(3), 773–780. <https://doi.org/10.30630/joiv.7.3.1035>

Lu, F., Ferraro, F., & Raff, E. (2022). *Continuously Generalized Ordinal Regression for Linear and Deep Models* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2202.07005>

Nedashkivskyi, B. (2024). HIERARCHICAL MULTI-LABEL CLASSIFICATION. *Věda a Perspektivy*, (5(36)). [https://doi.org/10.52058/2695-1592-2024-5\(36\)-347-356](https://doi.org/10.52058/2695-1592-2024-5(36)-347-356)

Nuryani, Hajairin, & Ilham. (2025). Analisis Yuridis Pemenuhan Unsur Tindak Pidana Penghinaan dan Ujaran Kebencian (Hate Speech) Perspektif KUHP dan UU ITE. *As-Syar i: Jurnal Bimbingan & Konseling Keluarga*, 7(4), 32–46. <https://doi.org/10.47467/as.v7i4.8437>

okkyibrohim. (n.d.). *Id-multi-label-hate-speech-and-abusive-language-detection/re_dataset.csv at master · okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection*. GitHub. Retrieved October 1, 2025, from https://github.com/okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection/blob/master/re_dataset.csv

Prabowo, F. A., Ibrohim, M. O., & Budi, I. (2019). Hierarchical Multi-label Classification to Identify Hate Speech and Abusive Language on Indonesian Twitter. *2019 6th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)*, 1–5. <https://doi.org/10.1109/ICITACEE.2019.8904425>

- Riadi, I., & Fadlil, A. (2023). *Analysis on Indonesian Twitter Utilizing*. 9(2).
- Rizky Pratama Putra Karo Karo. (2023). Hate Speech: Penyimpangan terhadap UU ITE, Kebebasan Berpendapat dan Nilai-Nilai Keadilan Bermartabat. *Jurnal Lemhannas RI*, 10(4), 52–65. <https://doi.org/10.55960/jlri.v10i4.370>
- Rokhim, I. F., Sarno, R., Septiyanto, A. F., Haryono, A. T., & Sabilla, S. I. (2024). IndoBERT-Based Ensemble Learning for Multi-Level Multi-Label Hate Speech Detection in Indonesian Social Media. *2024 Beyond Technology Summit on Informatics International Conference (BTS-I2C)*, 456–461. <https://doi.org/10.1109/BTS-I2C63534.2024.10942204>
- Sachdeva, P., Barreto, R., Bacon, G., & Sahn, A. (2022). *The Measuring Hate Speech Corpus: Leveraging Rasch Measurement Theory for Data Perspectivism*.
- Sari, T. I., Ardilla, Z. N., Hayatin, N., & Maskat, R. (2022). Abusive comment identification on Indonesian social media data using hybrid deep learning. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 11(3), 895. <https://doi.org/10.11591/ijai.v11.i3.pp895-904>
- Shaw, C., LaCasse, P., & Champagne, L. (2025). Exploring emotion classification of indonesian tweets using large scale transfer learning via IndoBERT. *Social Network Analysis and Mining*, 15(1), 22. <https://doi.org/10.1007/s13278-025-01439-6>
- Shi, X., Cao, W., & Raschka, S. (2023). Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. *Pattern Analysis and Applications*, 26(3), 941–955. <https://doi.org/10.1007/s10044-023-01181-9>
- Singh, N. K., Ghosh, K., Mahapatra, J., Garain, U., & Senapati, A. (n.d.). *HCDIR: End-to-end Hate Context Detection, and Intensity Reduction model for online comments*.
- Social, W. A. (2024, January 31). Digital 2024. *We Are Social Indonesia*. <https://wearesocial.com/id/blog/2024/01/digital-2024/>
- Susanto, L., Wijanarko, M. I., Pratama, P. A., Hong, T., Idris, I., Aji, A. F., & Wijaya, D. (2024). *IndoToxic2024: A Demographically-Enriched Dataset of Hate Speech and Toxicity Types for Indonesian Language* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2406.19349>
- Syafruddin, Thaba, A., & Ananda, R. (2024). UJARAN KEBENCIAN NETIZEN INDONESIA PADA AKUN TWITTER ES TEH: TINJAUAN LINGUISTIK FORENSIK. *Semantik*, 13(1), 15–28. <https://doi.org/10.22460/semantik.v13i1.p15-28>