ANALISIS SENTIMEN KABUR AJA DULU PADA KOMENTAR *YOUTUBE* MENGGUNAKAN *LARGE LANGUAGE MODEL*

TUGAS AKHIR

Diajukan Oleh :

NURUL MIRDAWATI
NIM. 220705027

Mahasiswa Fakultas Sains dan Teknologi

Program Studi Teknologi Informasi

FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI AR-RANIRY
BANDA ACEH
2026 M/1448 H

LEMBAR PERSETUJUAN

ANALISIS SENTIMEN KABUR AJA DULU PADA KOMENTAR *YOUTUBE* MENGGUNAKAN *LARGE LANGUAGE MODEL*

Tugas Akhir

Diajukan kepada Fakultas Sains dan Teknologi
Universitas Islam Negeri (UIN) Ar-Raniry Banda Aceh
Sebagai Salah Satu Beban Studi Memperoleh Gelar Sarjana (S1)
Dalam Ilmu Teknologi Informasi

Oleh:

Nurul Mirdawati
NIM. 220705027

Mahasiswa Fakultas Sains dan Teknologi
Program Studi Teknologi Informasi

Disetujui untuk Dimunaqasyahkan Oleh:

Pembimbing I,	Pembimbing II,
Dr. Hendri Ahmadian, S.Si., M.I.M	**Mulkan Fadhli, S.T, M.T**
NIP.198301042014031002	**NIP.198811282020121006**

Mengetahui,
Ketua Program Studi Teknologi Informasi

Malahayati, M.T
NIP. 198301272015032003

LEMBAR PENGESAHAN

ANALISIS SENTIMEN KABUR AJA DULU PADA KOMENTAR YOUTUBE MENGGUNAKAN LARGE LANGUAGE MODEL

TUGAS AKHIR

Telah Diuji oleh Panitia Ujian Munaqasyah Tugas Akhir
Fakultas Sains dan Teknologi UIN Ar-Raniry Banda Aceh dan Dinyatakan Lulus
Serta Diterima Sebagai Salah Satu Beban Studi Program Sarjana (S1)
Dalam Program Studi Teknologi Informasi

Pada Hari/Tanggal: Selasa, 18 Agustus 2026 M
4 Rabiul Awal 1448 H
Di Darussalam, Banda Aceh

Panitia Ujian Munaqasyah Tugas Akhir:

Ketua,	Sekretaris,
Dr. Hendri Ahmadian, M.I.M.	**Mulkan Fadhli, S.T., M.T.**
NIP. 198301042014031002	**NIP. 198811282020121006**
Penguji I,	Penguji II,
Khairan AR, M.Kom.	**Mahanavati, M.T.**
NIP. 198607042014031001	**NIP. 198301272015032003**

Mengetahui:
Dekan Fakultas Sains dan Teknologi
UIN Ar-Raniry Banda Aceh

Prof. Dr. Ir. Muhammad Dirhamsyah, M.T., IPU
NIP. 196210021988111001

LEMBAR PERNYATAAN KEASLIAN TUGAS AKHIR

Yang bertanda tangan di bawah ini:

Nama	:	Nurul Mirdawati
NIM	:	220705027
Program Studi	:	Teknologi Informasi
Fakultas	:	Sains dan Teknologi
Judul	:	Analisis Sentimen Kabur Aja Dulu Pada Komentar *Youtube* Menggunakan *Large Language Model*

Dengan ini menyatakan bahwa dalam penulisan tugas akhir ini, saya:

1. Tidak menggunakan ide orang lain tanpa mampu mengembangkan dan mempertanggungjawabkan;
2. Tidak melakukan plagiasi terhadap naskah karya orang lain;
3. Tidak menggunakan karya orang lain tanpa menyebutkan sumber asli atau tanpa izin pemilik karya;
4. Tidak memanipulasi dan memalsukan data;
5. Mengerjakan sendiri karya ini dan mampu bertanggungjawab atas karya ini.

Bila dikemudian hari ada tuntutan dari pihak lain atas karya saya, dan telah melalui pembuktian yang dapat dipertanggungjawabkan dan ternyata memang ditemukan bukti bahwa saya telah melanggar pernyataan ini, maka saya siap dikenai sanksi berdasarkan aturan yang berlaku di Fakultas Sains dan Teknologi UIN Ar-Raniry Banda Aceh. Demikian pernyataan ini saya buat dengan sesungguhnya dan tanpa paksaan dari pihak manapun.

Banda Aceh, 18 Agustus 2026

Yang Menyatakan,

Nurul Mirdawati

ABSTRAK

Fenomena Kabur Aja Dulu mencerminkan keresahan masyarakat Indonesia terhadap kondisi sosial ekonomi di platform *YouTube*. Penelitian ini bertujuan membandingkan performa model *DeepSeek*, *Gemma*, dan *IndoBERT* dalam mengklasifikasikan sentimen komentar terkait topik tersebut menggunakan metodologi *CRISP-DM*. Sebanyak 10.000 data diseimbangkan, diproses, dan dilabeli secara otomatis menggunakan leksikon *InSet*. *IndoBERT* dilatih dengan *full fine-tuning*, sementara *DeepSeek* dan *Gemma* diadaptasi menggunakan teknik *QLoRA*. Hasil evaluasi menunjukkan *IndoBERT* memberikan performa terbaik dengan akurasi 82,73% dan *Macro F1-score* 0,7913, serta paling tangguh dalam menangani teks informal dan ketidakseimbangan kelas. *DeepSeek* sangat sensitif pada kelas negatif (*recall* 0,9140) namun lemah pada kelas positif, sedangkan *Gemma* menunjukkan performa menengah. Sebagai model paling optimal, *IndoBERT* kemudian diimplementasikan ke dalam aplikasi web berbasis *Streamlit* untuk klasifikasi sentimen.

Kata Kunci: Analisis Sentimen, *YouTube*, Kabur Aja Dulu, *Large Language Model*, *IndoBERT*, *DeepSeek*, *Gemma*.

ABSTRACT

The Kabur Aja Dulu phenomenon reflects Indonesian public unrest regarding socio-economic conditions expressed on YouTube. This study aims to compare the performance of DeepSeek, Gemma, and IndoBERT models in classifying sentiments of related comments using the CRISP-DM methodology. A balanced dataset of 10,000 comments was pre-processed and pseudo-labeled using the InSet lexicon. IndoBERT was trained via full fine-tuning, while DeepSeek and Gemma used the QLoRA technique. Evaluation results indicate that IndoBERT achieved the best overall performance with 82.73% accuracy and a 0.7913 Macro F1-score, proving most robust against informal text and class imbalance. DeepSeek is highly sensitive to the negative class (0.9140 recall) but struggles with the positive class, whereas Gemma shows moderate performance. As the most optimal model, IndoBERT was subsequently deployed into a Streamlit-based web application for sentiment classification.

Keywords: *Sentiment Analysis, YouTube, Kabur Aja Dulu, Large Language Model, IndoBERT, DeepSeek, Gemma.*

KATA PENGANTAR

Bismillahirrahmanirrahim.

Puji syukur ke hadirat Allah SWT atas segala rahmat, hidayah, dan karunia-Nya sehingga penulis dapat menyelesaikan tugas akhir yang berjudul "Analisis Sentimen Kabur Aja Dulu Pada Komentar *Youtube* Menggunakan *Large Language Model*." Tak lupa pula Shalawat beriringkan salam tercurahkan kepada sang baginda Nabi Muhammad SAW. Tugas akhir ini disusun sebagai salah satu persyaratan untuk menyelesaikan pendidikan pada Program Studi Teknologi Informasi, Fakultas Sains dan Teknologi, Universitas Islam Negeri (UIN) Ar-Raniry Banda Aceh.

Penulis menyadari bahwa penyusunan tugas akhir ini masih memiliki banyak kekurangan. Oleh karena itu, kritik dan saran yang bersifat membangun sangat diharapkan untuk penyempurnaan di masa mendatang. Penyelesaian tugas akhir ini tidak terlepas dari dukungan dan bantuan berbagai pihak. Dengan penuh rasa terima kasih, penulis menyampaikan apresiasi kepada semua pihak yang telah membantu, baik secara langsung maupun tidak langsung, sejak proses penelitian hingga penyusunan laporan ini. Oleh karena itu, penulis ingin mengucapkan terima kasih dengan teramat dalam kepada:

1. Almarhum Ayah tercinta**,** meskipun sudah setahun ini Ayah tidak lagi menemani setiap langkah penulis, kasih sayang, doa, dan segala pengorbanan Ayah tetap hidup dalam hati dan menjadi kekuatan bagi penulis hingga hari ini. Semoga Allah SWT menempatkan Ayah di tempat terbaik di sisi-Nya. Mamak tersayang**,** terima kasih telah menjadi sumber kekuatan, doa, dan semangat yang tak pernah berhenti, serta selalu menemani penulis hingga mampu menyelesaikan Tugas Akhir ini. Karya ini penulis persembahkan dengan penuh cinta untuk Ayah dan Mamak.
2. Ibu Malahayati, M.T, selaku Ketua Program Studi Teknologi Informasi, yang telah banyak memberikan bimbingan serta pembelajaran kepada penulis.
3. Bapak Dr. Hendri Ahmadian, M.I.M. dan Bapak Mulkan Fadhli, S.T, M.T selaku dosen pembimbing 1 dan dosen pembimbing 2, di mana sudah memberikan arahan dan bimbingan selama proses pembuatan dan penyusunan tugas akhir penulis.

4. Ibu Cut Ida Rahmadiana, S.Si, selaku seseorang yang telah banyak membantu selama masa perkuliahan penulis.
5. Ucapan terima kasih yang teramat dalam kepada seseorang yang telah hadir dan banyak membantu penulis dalam setiap proses yang dilalui. Terima kasih atas segala rasa, kasih sayang, perhatian, dukungan, usaha, serta doa-doa baik yang selalu dipanjatkan untuk penulis. Semoga jalinan yang telah terjalin ini dapat terus berlayar, menemani setiap langkah hingga kapan pun.
6. Dea, Naqqa, dan Ojak, terima kasih telah menjadi teman yang selalu hadir dalam berbagai cerita selama masa perkuliahan. Terima kasih untuk setiap bantuan, candaan, semangat, dan kebersamaan yang membuat perjalanan ini terasa lebih ringan. Semoga apa pun jalan yang kita pilih setelah ini, kita tetap menjadi sahabat yang saling mengingat dan mendoakan.
7. Teman-teman Program Studi Teknologi Informasi angkatan 2022 dan teman-teman terdekat yang telah menjadi bagian dari perjalanan penulis selama masa perkuliahan. Terima kasih atas kebersamaan, bantuan, dukungan, serta semangat yang diberikan selama menjalani masa studi hingga proses penyelesaian Tugas Akhir ini.
8. Terakhir, terima kasih kepada diri sendiri, Nurul Mirdawati, karena telah memilih untuk tetap bertahan dan tidak menyerah maupun berhenti di tengah perjalanan. Terima kasih telah melewati berbagai rasa lelah, kecewa, takut, dan ragu, namun tetap berusaha melangkah meskipun tidak selalu mudah.

Banda Aceh, 18 Agustus 2026

Penulis,

Nurul Mirdawati

NIM. 220705027

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I
PENDAHULUAN

1.1 Latar Belakang

Media sosial telah mengalami perkembangan yang pesat dalam beberapa tahun terakhir dan menjadi bagian dari kehidupan masyarakat sebagai wadah komunikasi dan interaksi digital. Media sosial digunakan oleh miliaran orang di seluruh dunia dan telah dengan cepat menjadi salah satu teknologi yang paling menentukan di era kita (Appel et al., 2020). Salah satu platform media sosial yang telah mengalami pertumbuhan signifikan adalah *YouTube*. Pengguna dapat mengunggah video, mengirim pesan, dan berbagi pendapat dengan orang lain menggunakan platform *YouTube* (Alhujaili & Yafooz, 2021). Setiap video yang diunggah sering kali disertai dengan ribuan komentar, yang bisa menjadi cermin dari opini dan perasaan publik terhadap suatu topik tertentu. Salah satu contohnya adalah topik ”Kabur Aja Dulu” yang muncul di tengah situasi sosial, ekonomi, dan politik Indonesia yang dinamis. Kondisi ini terjadi ketika masyarakat tertekan oleh keterbatasan lapangan pekerjaan, biaya pendidikan yang tinggi, serta ekonomi yang tidak stabil, sehingga mulai mempertimbangkan untuk mencari peluang hidup yang lebih baik di luar negeri. Tagar (#) "Kabur Aja Dulu" menjadi simbol kekecewaan terhadap kondisi ekonomi, sosial, dan kebijakan pemerintah yang dinilai kurang berpihak pada masyarakat (Sari et al., 2025). Platform *YouTube* menjadi salah satu media bagi masyarakat untuk mengungkapkan pendapat terkait topik “Kabur Aja Dulu” yang dituangkan dalam komentar-komentar dengan beragam sentimen, baik yang bernada negatif maupun positi

Analisis sentimen merupakan alat yang kuat bagi pelaku bisnis, pemerintah, dan peneliti untuk mengekstrak dan menganalisis suasana hati serta pandangan publik, guna mendapatkan wawasan bisnis, dan membuat keputusan yang lebih baik (Birjali et al., 2021). Pelaksanaan analisis sentimen terhadap komentar-komentar di *YouTube* pada topik “Kabur Aja Dulu” memiliki tujuan untuk mendapatkan wawasan mendalam mengenai persepsi masyarakat. Untuk dapat melakukan

analisis sentimen dalam komentar-komentar *YouTube*, teknologi *Natural Language Processing* digunakan sebagai alat utama. *NLP* adalah cabang dari kecerdasan buatan yang berfokus pada interaksi antara komputer dan bahasa manusia. Bidang *NLP*, yang juga dikenal sebagai linguistik komputasional, melibatkan rekayasa model dan proses komputasi untuk menyelesaikan masalah praktis dalam memahami bahasa manusia (Otter et al., 2021).

Berbagai penelitian terdahulu telah menerapkan beragam model untuk tugas ini di platform *YouTube*. Salah satu penelitian yang berfokus pada model machine learning klasik oleh (Shivsharan et al., 2025) mengevaluasi berbagai model dan menemukan bahwa *Bagging Classifier* memberikan akurasi terbaik, mencapai 95,82%. Senada dengan itu, (Angdresey et al., 2025) menerapkan *Multinomial Naïve Bayes* untuk menganalisis sentimen pada debat politik di *YouTube* Indonesia (sebuah konteks berbahasa Indonesia) dan berhasil mencapai akurasi 85,155%. Seiring berkembangnya arsitektur *Transformer*, penelitian beralih ke model yang lebih canggih. (Gadyanavar et al., 2025) mengimplementasikan *BERT* untuk mengklasifikasikan komentar ke dalam lima label sentimen (dari "Sangat Negatif" hingga "Sangat Positif") dan melaporkan akurasi exact match sebesar 67%. Penelitian lebih lanjut berfokus pada varian yang lebih optimal yaitu *RoBERTa*. (Yadalam et al., 2025) menggunakan *RoBERTa* untuk mengklasifikasikan sentimen kecemasan pada komentar kesehatan mulut dan mencapai akurasi 75,00%. Sementara itu, (El Azzouzy et al., 2025) membandingkan beberapa arsitektur *Transformer* (termasuk *BERT*, *GPT*, dan *T5*) dan menyimpulkan bahwa *RoBERTa* memberikan kinerja superior, dengan akurasi 95% dan F1-score 0,93.

Namun, penelitian-penelitian tersebut umumnya berfokus pada model *machine learning* klasik atau arsitektur *Transformer* generasi sebelumnya (seperti *BERT* dan *RoBERTa*). Sementara itu, terdapat model seperti *IndoBERT* dikembangkan khusus untuk Bahasa Indonesia. Selain itu, saat ini telah berkembang *Large Language Models* global generasi terbaru seperti *DeepSeek* dan *Gemma* yang diklaim memiliki pemahaman konteks yang lebih kuat. Secara khusus, perbandingan kinerja antara model *DeepSeek*, *Gemma* dan *IndoBERT* dalam menganalisis topik ”Kabur Aja Dulu” masih perlu dieksplorasi lebih lanjut.

Oleh karena itu, penelitian ini bertujuan untuk menganalisis perbedaan kemampuan model *DeepSeek*, *Gemma,* dan *IndoBERT* dalam mengklasifikasikan sentimen positif dan negatif pada komentar *YouTube* topik "Kabur Aja Dulu", serta menentukan model yang memberikan performa keseluruhan terbaik berdasarkan *confusion matrix* dan *classification report.* Hasil penelitian ini diharapkan dapat memberikan wawasan mengenai pola sentimen pada data yang dianalisis sekaligus menghasilkan tolok ukur empiris kinerja ketiga model dalam menangani Bahasa Indonesia yang dinamis di media sosial.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, penelitian ini dirumuskan dalam dua pertanyaan utama sebagai berikut:

1. Bagaimana perbedaan kemampuan model *DeepSeek*, *Gemma*, dan *IndoBERT* dalam mengklasifikasikan sentimen negatif dan positif pada komentar YouTube terkait topik "Kabur Aja Dulu"?
2. Model manakah yang memberikan performa keseluruhan terbaik dalam analisis sentimen komentar *YouTube* terkait topik "Kabur Aja Dulu"?

1.3 Tujuan Penelitian

Penelitian ini bertujuan sebagai berikut:

1. Menganalisis perbedaan kemampuan model *DeepSeek*, *Gemma*, dan *IndoBERT* dalam mengklasifikasikan sentimen negatif dan positif pada komentar YouTube terkait topik "Kabur Aja Dulu".
2. Menentukan model yang memberikan performa keseluruhan terbaik dalam analisis sentimen komentar *YouTube* terkait topik "Kabur Aja Dulu".

1.4 Manfaat Penelitian

Manfaat dari penelitian ini sebagai berikut:

1. Kontribusi pada Bidang *NLP* Bahasa Indonesia, penelitian ini menyajikan data kinerja empiris dari *Large Language Models DeepSeek*, *Gemma* dan model spesifik Bahasa Indonesia *IndoBERT*. Hasil ini berkontribusi pada

pemahaman tentang efektivitas arsitektur model baru dalam menangani teks Bahasa Indonesia di media sosial.

2. Memberikan Pemahaman tentang Topik Sosial, penelitian ini memberikan wawasan objektif berbasis data mengenai sentimen dan persepsi publik terhadap topik "Kabur Aja Dulu". Hasil analisis ini dapat membantu sosiolog, jurnalis, dan pemangku kepentingan dalam memahami dinamika sosial yang terjadi di masyarakat Indonesia melalui platform *YouTube*.
3. Menjadi Rujukan bagi Peneliti dan Praktisi, hasil evaluasi kinerja dari model *DeepSeek*, *Gemma*, dan *IndoBERT* yang dapat menjadi referensi metodologis bagi peneliti dalam memilih model yang paling sesuai untuk tugas analisis sentimen pada media sosial berbahasa Indonesia.

1.5 Batasan Masalah

Untuk menjaga fokus penelitian agar terarah dan dapat diselesaikan secara efektif, penelitian ini dibatasi pada hal-hal berikut:

1. Data penelitian berasal dari komentar *YouTube* pada video yang berkaitan dengan topik "Kabur Aja Dulu".
2. Model yang digunakan terbatas pada *DeepSeek*, *Gemma*, dan *IndoBERT*.
3. Analisis hanya mengklasifikasikan sentimen positif dan negatif berdasarkan leksikon *InSet* sebagai *pseudo-label*, tanpa menilai sikap terhadap topik dan tanpa anotasi manusia.
4. Dataset dibagi menjadi 70% data latih, 15% data validasi, dan 15% data uji.
5. Kinerja setiap model dievaluasi menggunakan *confusion matrix* untuk melihat distribusi hasil prediksi dan *classification report* untuk mengukur *precision*, *recall*, *F1-score*, serta *accuracy.*

1.6 Metodologi Penelitian

Penelitian ini menggunakan jenis penelitian eksperimental kuantitatif dengan mengadaptasi kerangka kerja *Cross-Industry Standard Process for Data Mining*. Tahapan penelitian diawali dengan fase pemahaman masalah untuk mendefinisikan tujuan analisis sentimen pada topik Kabur Aja Dulu. Selanjutnya pada tahap pemahaman data pengumpulan komentar *YouTube* dilakukan secara otomatis menggunakan alat *Communalytics*. Fase persiapan data merupakan proses yang

mencakup pembersihan teks, normalisasi bahasa gaul, stemming, pelabelan sentimen otomatis berbasis kamus leksikon *InSet*, dan penyeimbangan data hingga menghasilkan 10.000 dataset final. Data tersebut kemudian dibagi menjadi 70% data latih, 15% data validasi, dan 15% data uji. Pada tahap pemodelan tiga model diimplementasikan secara terprogram. Kinerja dari ketiga model tersebut selanjutnya diukur pada tahap evaluasi menggunakan metrik *confusion matrix* dan *classification report*. Tahap terakhir adalah penyajian hasil di mana model paling optimal diimplementasikan ke dalam aplikasi web berbasis *Streamlit*

1.7 Sistematika Penulisan

Adapun sistematika penulisan laporan penelitian ini disusun secara terstruktur dalam lima bab utama berikut:

BAB I PENDAHULUAN

Bagian ini bertujuan untuk memberikan gambaran umum tentang latar belakang, rumusan masalah, tujuan penelitian, batasan masalah, manfaat penelitian, dan sistematika penelitian.

BAB II TINJAUAN PUSTAKA

Tinjauan pustaka menyediakan dasar teori dan penelitian terdahulu yang relevan dengan topik.

BAB III METODOLOGI PENELITIAN

Metodologi penelitian menggambarkan cara penelitian akan dilakukan. Penjelasan tentang jenis penelitian, tahapan penelitian, subjek dan objek, teknik pengumpulan data, dan instrumen lainnya.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menjelaskan mengenai tahapan dan hasil dari pengujian yang dilakukan dari tiap-tiap metode yang diuji.

BAB V KESIMPULAN DAN SARAN

Bab ini merupakan penutup yang berisi kesimpulan dari hasil penelitian yang dilakukan dan saran untuk penelitian selanjutnya

BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian mengenai analisis sentimen pada komentar *YouTube* telah berevolusi dari penggunaan algoritma tradisional hingga model bahasa berbasis kecerdasan buatan. Pada tahap awal, pendekatan berfokus pada model *Machine Learning* klasik karena kecepatan komputasi dan kemudahan implementasinya. Sebagai contoh, penelitian oleh (Shivsharan et al., 2025) mengevaluasi berbagai algoritma *ML* dan menemukan bahwa *Bagging Classifier* mampu mencapai akurasi 95,82%. Dalam konteks berbahasa Indonesia, (Angdresey et al., 2025)menerapkan pendekatan hibrida dengan menjadikan *Multinomial Naive Bayes* sebagai inti klasifikasi sentimen debat politik. Meskipun model-model klasik ini mampu menghasilkan akurasi komputasi yang memadai, pendekatan ini memiliki keterbatasan mendasar. Algoritma *ML* tradisional umumnya sangat bergantung pada rekayasa fitur manual dan kerap kesulitan dalam menangkap nuansa bahasa yang tidak terstruktur, sarkasme, serta konteks kalimat yang kompleks, sehingga mendorong transisi penelitian menuju arsitektur deep learning.

Untuk mengatasi keterbatasan pemahaman konteks tersebut, tren penelitian beralih pada arsitektur *Transformer* yang mampu menangkap relasi antar kata secara *bidireksional*. Penelitian oleh (Gadyanavar et al., 2025) mengimplementasikan *BERT* multilingual pada komentar *YouTube* dan mencapai akurasi 67%. Mengingat ruang peningkatan yang masih terbuka, studi selanjutnya beralih pada varian *Transformer* yang lebih optimal seperti *RoBERTa*. (Yadalam et al., 2025) menggunakan *RoBERTa* untuk mendeteksi sentimen kecemasan pada komentar kesehatan mulut dengan akurasi 75,00%, sementara (El Azzouzy et al., 2025) menyimpulkan bahwa *RoBERTa* secara konsisten mengungguli *BERT*, *GPT*, dan *T5* dengan akurasi 95% pada teks berbahasa Inggris. Walaupun arsitektur Transformer generasi awal ini menawarkan peningkatan pemahaman semantik yang signifikan, model-model global tersebut sering kali kurang optimal ketika

dihadapkan pada dialek internet yang sangat informal, penuh singkatan, dan ragam bahasa gaul yang spesifik pada suatu negara.

Merespons tantangan lokalisasi bahasa, penerapan model *Transformer* yang dilatih khusus untuk Bahasa Indonesia mulai berkembang pesat, dengan *IndoBERT* sebagai salah satu model yang paling mendominasi. (Aulia Ruslani & Fauzan, 2025) membuktikan efektivitas *IndoBERT* pada ulasan *e-commerce* dengan akurasi 94,6%. Keunggulan model *Transformer* lokal dalam memproses opini publik terhadap isu sosial dan kebijakan juga dikonfirmasi oleh berbagai studi. *IndoBERT* terbukti superior dibandingkan metode pelabelan leksikon manual pada kasus topik perpajakan Coretax (Rizkia et al., 2025), efektif dalam mengklasifikasikan sentimen kenaikan PPN 12% di lintas platform media sosial (Manoppo et al., 2025), serta tangguh dalam mengurai persepsi publik terhadap ibu kota Nusantara di platform media sosial berbasis teks singkat (Salma et al., 2025). Selain arsitektur *encoder*, arsitektur *Text-to-Text Transfer Transformer* (*T5*) juga mulai diadopsi secara lokal, seperti yang ditunjukkan oleh (Suhendar et al., 2025) pada data transkripsi audio.

Serangkaian penelitian tersebut menegaskan kapabilitas *IndoBERT* sebagai fondasi yang solid untuk pemrosesan teks Bahasa Indonesia. Namun, seiring dengan pesatnya perkembangan kecerdasan buatan, saat ini telah bermunculan *Large Language Models* generatif berukuran masif seperti *DeepSeek* dan *Gemma*. Model-model global generasi terbaru ini diklaim memiliki kemampuan penalaran dan pemahaman konteks multi-bahasa yang jauh melebihi generasi sebelumnya. Meskipun demikian, sebuah celah penelitian (*research gap*) yang krusial masih terbuka, di mana belum ada studi yang secara empiris membandingkan kinerja *DeepSeek*, *Gemma* dan *IndoBERT* secara langsung pada studi kasus komentar *YouTube* berbahasa Indonesia yang sangat informal dan dinamis, seperti pada isu sosial "Kabur Aja Dulu". Rangkuman komparatif dari seluruh penelitian terdahulu yang menyoroti celah tersebut dan menjadi landasan bagi penelitian ini disajikan secara lengkap pada Tabel 2.1.

Tabel 2. 1 Penelitian Terdahulu

Judul Penelitian (Tahun)	Hasil Penelitian	Persamaan	Perbedaan
Evaluating Machine Learning Models for Sentiment Analysis of YouTube Comments and Creating an Accessible Web Application for Comment Analysis (Shivsharan et al., 2025)	Mengevaluasi enam model machine learning klasik (*XGBoost*, *RF*, *LR*, *GDBT*, *AdaB*, *BgC*) dan menemukan *Bagging Classifier* memiliki akurasi tertinggi, yaitu 95,82%.	Menganalisis sentimen dari komentar *YouTube* dan memiliki tujuan untuk mengevaluasi kinerja beberapa model.	Fokus pada model machine learning klasik, bukan LLM modern *DeepSeek*, *Gemma*, dan *IndoBERT* yang dilakukan dalam penelitian ini.
Sentiment Analysis for Political Debates on YouTube Comments using BERT Labeling, Random Oversampling, and Multinomial Naïve Bayes (Angdresey et al., 2025)	Model klasifikasi utamanya adalah *Multinomial Naïve Bayes* yang dikombinasikan dengan *Random Oversampling*. Data dilabeli secara otomatis menggunakan "*Indonesian BERT Base Sentiment Classifier*". Model *MNB* yang dihasilkan mencapai akurasi 85,155%.	Menganalisis sentimen pada komentar *YouTube* berbahasa Indonesia terkait topik sosial-politik.	Model utama adalah *ML* klasik, berbeda dengan penelitian ini yang berfokus pada LLM. Dalam studi tersebut, *IndoBERT* hanya dilakukan untuk labeling dan tidak dievaluasi kinerjanya, sementara penelitian ini secara khusus mengevaluasi kinerja *DeepSeek*, *Gemma*, dan *IndoBERT*.
Sentiment Analysis of YouTube Comments Using Bidirectional Encoder Representations	Mengklasifikasikan komentar *YouTube* ke dalam lima label sentimen (1: "Sangat Negatif" hingga 5: "Sangat Positif")	Menganalisis sentimen dari komentar *YouTube* dan menggunakan model berbasis	Menggunakan *BERT* multilingual umum, bukan LLM baru *DeepSeek*, *Gemma* dan

from Transformers Neural Network Model (Gadyanavar et al., 2025)	menggunakan model "*Bert-base-multilingual-uncased-sentiment*". Model ini dilaporkan mencapai akurasi sebesar 67%.	arsitektur Transformer.	*IndoBERT* yang menjadi fokus penelitian ini.
An explainable ROBERTa approach to analyzing panic and anxiety sentiment in oral health education YouTube comments (Yadalam et al., 2025)	Model *RoBERTa* mencapai akurasi 75,00% dalam mengklasifikasikan komentar YouTube (terkait kesehatan mulut) ke dalam empat kategori: *confusing*, *informative*, *panic*, dan *unproductive*.	Menggunakan komentar *YouTube* sebagai data untuk analisis sentimen dan menerapkan model berbasis arsitektur Transformer.	Menggunakan *RoBERTa*, berbeda dengan penelitian ini yang berfokus pada LLM generasi baru *DeepSeek*, *Gemma* dan *IndoBERT*.
Transformer-based models for sentiment analysis of YouTube video comments (El Azzouzy et al., 2025)	Melakukan studi komparatif antara *BERT*, *GPT*, *RoBERTa*, dan *T5*. Ditemukan bahwa *RoBERTa* memberikan kinerja superior dengan akurasi 95%.	Melakukan evaluasi model-model *Transformer*, dan juga menganalisis sentimen dari komentar *YouTube*.	Membandingkan Transformer generasi awal (*BERT*, *GPT*, *RoBERTa*, dan *T5*) pada data berbahasa Inggris. Berbeda dengan fokus penelitian ini pada LLM baru *DeepSeek*, *Gemma* dan *IndoBERT* untuk Bahasa Indonesia.
Implementasi Model *Transformer* untuk Analisis Sentimen dan Ringkasan Ulasan *E-commerce* Berbahasa Indonesia (Aulia	Menggunakan *IndoBERT* (akurasi 94,6%) untuk sentimen dan *IndoT5* untuk ringkasan ulasan produk.	Menggunakan model *Transformer* untuk analisis sentimen Bahasa Indonesia.	Objek penelitian adalah ulasan produk *e-commerce*, bukan komentar *YouTube*.

Ruslani & Fauzan, 2025)			
Analisis Sentimen Berbasis *Transformer*: Persepsi Publik terhadap Nusantara pada Perayaan Kemerdekaan Indonesia yang Pertama (Salma et al., 2025)	Model *TextBlob-IndoBERT* memberikan hasil terbaik pada teks informal *Twitter* terkait perayaan kemerdekaan.	Menggunakan model *Transformer* (*IndoBERT*) untuk analisis sentimen Bahasa Indonesia.	Studi kasus pada *Twitter*, sedangkan penelitian ini pada *YouTube*.
Analisis Sentimen Hasil Transkripsi Audio Berbahasa Indonesia Menggunakan *T5* (*Text-to-Text Transfer Transformer*) (Suhendar et al., 2025)	Menggunakan model *T5* untuk analisis sentimen transkripsi audio. Mencapai akurasi 83% dan *F1-score* 0,83.	Menggunakan model *Transformer* untuk analisis sentimen.	Input data berasal dari transkripsi audio, sedangkan penelitian ini fokus pada teks komentar tertulis.
Analisis Sentimen Publik di Media Sosial terkait Rencana Kenaikan PPN 12% Menggunakan *IndoBERT* (Manoppo et al., 2025)	Menggunakan *IndoBERT* pada data X, Instagram, dan TikTok. Mencapai akurasi 84,94%.	Menggunakan model Transformer untuk analisis sentimen Bahasa Indonesia.	Fokus pada kebijakan publik (PPN 12%) di multi-platform media sosial, bukan spesifik *YouTube*.
Analisis Sentimen Coretax: Perbandingan Pelabelan Data *Manual*, *Transformers-Based*, dan *Lexicon-Based* pada Performa *IndoBERT* (Rizkia et al., 2025)	Membandingkan beberapa pendekatan pelabelan data sentimen untuk topik Coretax. Hasil menunjukkan bahwa *IndoBERT* dan *IndoBERTweet*	Menggunakan berbagai varian model Transformer untuk analisis sentimen.	Sentimen terhadap topik Coretax (pajak/kebijakan) di media sosial, bukan komentar *YouTube*.

	memberikan kinerja lebih baik dibanding metode berbasis aturan.		

2.2 Kajian Teori

2.2.1 Media Sosial

Media sosial merupakan platform digital yang memungkinkan pengguna berinteraksi, berbagi informasi, dan mengekspresikan opini secara *online*. Platform ini telah mengubah lanskap komunikasi, termasuk komunikasi politik, dengan membentuk kembali cara figur publik berinteraksi dengan masyarakat dan mengungkapkan pandangan mereka (Gosztonyi et al., 2025). *YouTube* adalah salah satu platform media sosial terbesar dan salah satu repositori konten video terbesar, tidak hanya menyajikan konten video tetapi juga menyediakan ruang interaksi melalui jutaan komentar pengguna setiap hari, yang mencerminkan beragam opini, emosi, dan sentimen (Krishnappa K et al., 2025). Komentar-komentar yang muncul di platform *YouTube* mencerminkan spektrum emosi masyarakat terhadap isu yang sedang berkembang (Kusoema & Ibrahim, 2025), serta menjadi cermin pandangan publik terhadap berbagai topik sosial, ekonomi, dan politik.

2.2.2 *Natural Language Processing*

Natural Language Processing adalah salah satu cabang ilmu Kecerdasan Buatan yang mempelajari komunikasi antara manusia dan komputer melalui bahasa alami (Puspitasari et al., 2024). Tujuan utama *NLP* adalah pemahaman otomatis terhadap bahasa semi-terstruktur yang digunakan manusia (Ranjan et al., 2016). Jadi bukan sekadar memproses kata per kata, tetapi juga mencakup kemampuan untuk mengekstrak makna, memahami konteks, mengidentifikasi niat, dan bahkan mengenali sentimen di balik teks. Secara historis, *NLP* memiliki sejarah yang kaya yang dimulai sejak tahun 1950-an, diawali dengan pekerjaan awal pada penerjemahan mesin dan teori linguistik (Supriyono et al., 2024). Kini, *NLP* modern didominasi oleh pendekatan statistik, machine learning, dan deep learning. Model-model ini mampu "belajar" pola dan nuansa bahasa langsung dari data teks dalam jumlah besar. Evolusi inilah yang memungkinkan terciptanya aplikasi-aplikasi

canggih yang kita gunakan sehari-hari, seperti analisis sentimen pada ulasan produk atau komentar media sosial, chatbot yang responsif, asisten virtual (seperti *Google Assistant* atau *Siri*), peringkasan dokumen otomatis, dan deteksi berita palsu.

2.2.3 Analisis Sentimen

Salah satu tugas klasifikasi yang paling penting dalam studi *NLP* adalah analisis sentimen (Lin & Nuha, 2023). Analisis sentimen adalah teknik untuk mengidentifikasi dan mengklasifikasikan opini atau perasaan yang terkandung dalam teks menjadi kategori sentimen, misalnya positif atau negatif. Pada tingkat yang lebih mendalam, analisis sentimen dilakukan untuk mendeteksi dan mengukur ekspresi emosional dalam komentar, termasuk emosi seperti marah, antisipasi, jijik, gembira, takut, sedih, terkejut, dan percaya (Ma'aly et al., 2024).

2.2.4 *Large Language Model*

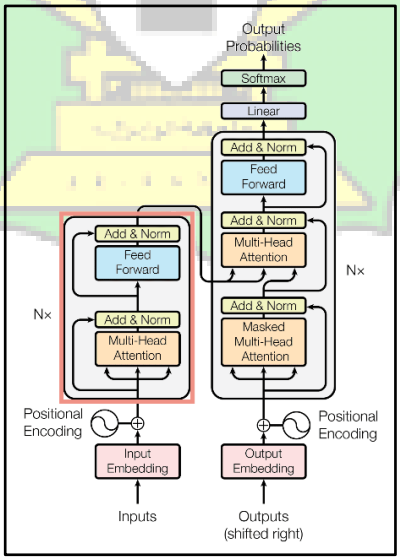
Large Language Model adalah kelas model Kecerdasan Buatan yang dilatih pada volume data teks yang sangat besar. Model-model ini dirancang untuk memahami dan menghasilkan bahasa manusia dengan tingkat akurasi dan konteks yang luar biasa. LLM pada dasarnya adalah sistem prediktif yang belajar untuk mengidentifikasi pola-pola kompleks dalam urutan kata, yang memungkinkan mereka melakukan berbagai tugas *Natural Language Processing* seperti penerjemahan, ringkasan, penjawaban pertanyaan, dan analisis sentimen. Aplikasi seperti *ChatGPT*, *Gemini*, dan lainnya yang berbasis pada *Large Language Models* menawarkan kemampuan analisis teks yang canggih, memungkinkan pengguna untuk mengidentifikasi informasi utama dengan cepat dan efisien (Yudertha & Dwimalida Putri, 2025).

2.2.5 *Transformer*

Transformer adalah arsitektur *deep learning* yang diperkenalkan oleh (Vaswani et al., 2017) dalam artikel jurnal yang berjudul "*Attention Is All You Need*". Arsitektur ini merevolusi bidang *NLP* dengan memperkenalkan mekanisme *self-attention*, yang memungkinkan model untuk menimbang pentingnya kata-kata yang berbeda dalam sebuah kalimat secara bersamaan, tanpa bergantung pada urutan sekuensial. Dalam *Natural Language Processing*, model berbasis *transformer* dan teknik *deep learning* telah menunjukkan potensi besar dalam meningkatkan presisi

dan konsistensi berbagai aplikasi *NLP* (Supriyono et al., 2024). Baru-baru ini, arsitektur *Transformer* telah secara signifikan memajukan penelitian *NLP* dengan memanfaatkan model pre-trained (Ahmadian et al., 2023).

Secara arsitektural, *Transformer* menggunakan mekanisme *Self-Attention* yang beroperasi melalui blok *Encoder* dan *Decoder*. Aliran data dimulai dengan memetakan teks (*Inputs*) menjadi representasi vektor melalui *Input Embedding*, yang kemudian digabungkan dengan *Positional Encoding* untuk mempertahankan informasi urutan kata. Data ini masuk ke dalam blok *Encoder* yang memprosesnya melalui mekanisme *Multi-Head Attention* dan *Feed Forward*, di mana model mengenkode setiap input dengan menimbang tingkat kepentingan (*weight*) setiap kata terhadap kata lainnya untuk memahami konteks global secara utuh (Pandey & Jain, 2021). Selanjutnya, blok *Decoder* menggunakan representasi dari *Encoder* tersebut beserta input prediksi sebelumnya (*Outputs shifted right*) melalui struktur lapisan yang serupa, namun dilengkapi *Masked Multi-Head Attention* agar model tidak memprediksi berdasarkan informasi kata di masa depan. Pada tahap akhir, hasil komputasi berulang dari *Decoder* dilewatkan pada lapisan *Linear* dan fungsi aktivasi *Softmax* untuk menghasilkan distribusi probabilitas (*Output Probabilities*).

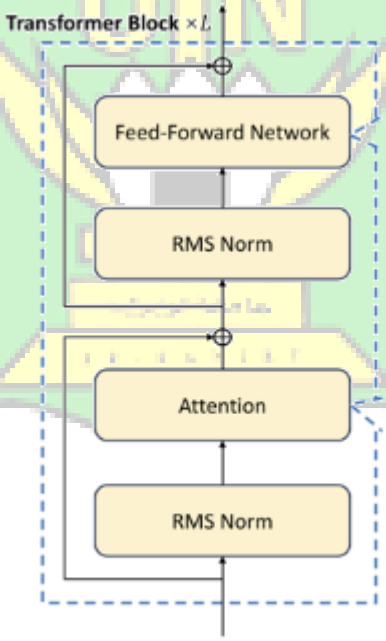


Gambar 2. 1 Arsitektur Transformer

2.2.6 Deepseek

DeepSeek merupakan keluarga model bahasa besar yang dikembangkan oleh *DeepSeek AI*. Penelitian ini secara khusus menggunakan *DeepSeek-R1-Distill-*

Qwen-1.5B, yaitu model dense dengan arsitektur *decoder-only* berbasis *Qwen2*. Model tersebut dikembangkan melalui proses distilasi berbasis data dengan menggunakan *Qwen2.5-Math-1.5B* sebagai model dasar dan data penalaran yang dihasilkan oleh *DeepSeek-R1* sebagai data pelatihan (DeepSeek-AI et al., 2025; Yang et al., 2024). Berbeda dengan *DeepSeek-R1* penuh yang menggunakan arsitektur *Mixture-of-Experts*, model hasil distilasi ini menggunakan arsitektur *dense*, sehingga seluruh parameter pada setiap lapisan digunakan dalam pemrosesan token. Sebagai *causal language model*, model memproses teks secara autoregresif dengan memprediksi token berikutnya berdasarkan rangkaian token yang telah diproses sebelumnya (Yang et al., 2024). Skala sekitar 1,5 miliar parameter membuat model ini jauh lebih ringan dibandingkan model *DeepSeek-R1* penuh dan lebih memungkinkan untuk diadaptasi menggunakan teknik *parameter-efficient fine-tuning*, seperti *Quantized Low-Rank Adaptation* atau *QLoRA* (Dettmers et al., 2023).



Gambar 2. 2 Arsitektur Decoder-Only pada DeepSeek

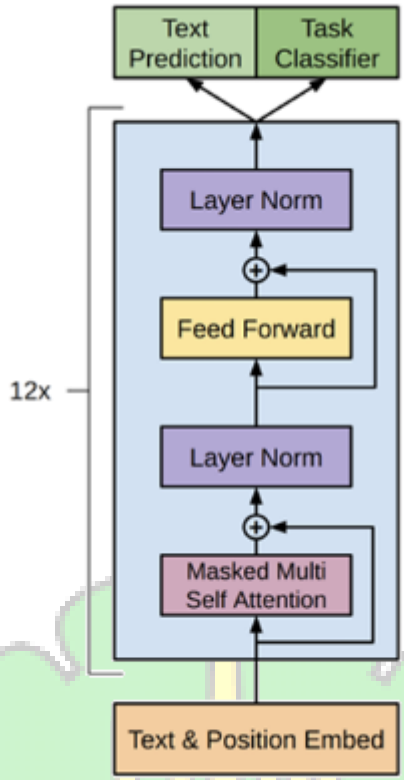
Secara arsitektural dan mekanis, seperti yang diilustrasikan pada Gambar 2.2, DeepSeek dibangun menggunakan tumpukan blok decoder. Aliran data dimulai dengan mengubah teks masukan menjadi representasi vektor (*input embedding*). Vektor ini kemudian diproses melalui lapisan-lapisan utama, di mana komponen

kuncinya adalah *masked self-attention*. Mekanisme bertopeng (*masked*) ini secara tegas memastikan bahwa pada setiap langkah komputasi matriks, model hanya dapat memperhatikan kata-kata (token) yang mendahuluinya dan mencegah model melihat token di masa depan. Setelah relasi antar kata diekstraksi, data diteruskan ke jaringan *feed-forward*. Proses komputasi pada blok decoder ini dilakukan secara iteratif dan autoregresif, di mana keluaran dari satu langkah menjadi masukan untuk langkah berikutnya, hingga model memprediksi probabilitas kata berikutnya secara berurutan untuk menghasilkan teks keluaran yang sesuai dengan instruksi awal (prompt).

2.2.7 *Gemma*

Gemma merupakan keluarga model bahasa dengan bobot terbuka dan berukuran relatif ringan yang dikembangkan oleh *Google* berdasarkan penelitian dan teknologi yang juga digunakan dalam pengembangan keluarga *Gemini* (Gemma Team, 2024). Penelitian ini menggunakan *gemma-1.1-2b-it*, yaitu pembaruan dari versi *instruction-tuned Gemma 2B* yang dikembangkan untuk meningkatkan kualitas respons, kemampuan mengikuti instruksi, faktualitas, pemrograman, dan percakapan multigilir.

Model tersebut merupakan *text-to-text causal language model* dengan arsitektur *Transformer decoder-only* yang bersifat *dense*. Model ini bukan model *multimodal* dan tidak menggunakan arsitektur *Mixture-of-Experts* (Gemma Team, 2024). Varian *gemma-1.1-2b-it* telah melalui proses *instruction tuning* dan penyempurnaan menggunakan metode *Reinforcement Learning from Human Feedback* untuk meningkatkan kemampuan mengikuti instruksi serta menghasilkan respons percakapan yang lebih berkualitas (Gemma Team, 2024). Dalam penelitian ini, model tersebut diadaptasi menggunakan *QLoRA* untuk melakukan klasifikasi sentimen berbasis prompt. Teknik *QLoRA* memungkinkan proses adaptasi dilakukan secara lebih efisien dengan mengombinasikan kuantisasi bobot model dan pelatihan adaptor berperingkat rendah (Dettmers et al., 2023).

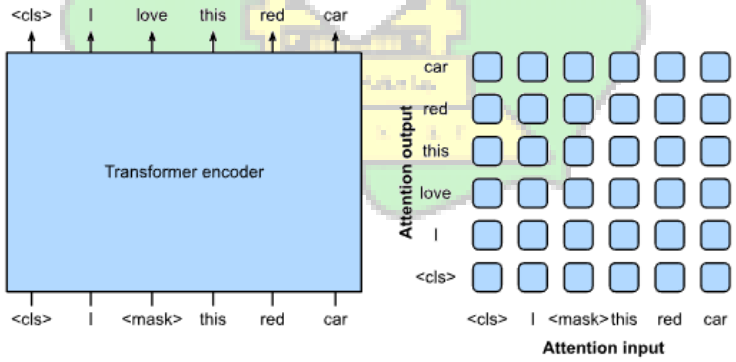


Gambar 2. 3 Arsitektur Decoder-Only pada Gemma

Merujuk pada Gambar 2.3, cara kerja Gemma didasarkan pada aliran data murni melalui blok *decoder Transformer* secara sekuensial tanpa intervensi blok *encoder*. Ketika menerima teks instruksi, lapisan awal memetakan teks tersebut menjadi representasi numerik dan menggabungkannya dengan informasi posisi kata. Data ini masuk ke dalam blok *decoder* yang memprosesnya menggunakan mekanisme *masked multi-head attention*. Melalui mekanisme ini, Gemma melakukan perhitungan bobot atensi untuk menimbang tingkat prioritas dan konteks setiap kata terhadap kalimat secara keseluruhan, dengan aturan ketat agar informasi dari token yang belum muncul tetap tersembunyi (*masked*). Hasil komputasi dari lapisan atensi kemudian diolah oleh jaringan saraf umpan maju (*feed-forward network*) yang padat. Berkat penyempurnaan instruksi yang telah dilewatinya, lapisan-lapisan internal *Gemma* secara presisi mengarahkan representasi kontekstual ini menuju lapisan linier terakhir agar probabilitas token yang dihasilkan sesuai dengan format klasifikasi yang diminta oleh pengguna

2.2.8 *IndoBERT*

IndoBERT adalah model bahasa yang secara khusus dilatih dan diadaptasi untuk bahasa Indonesia (Fatani & Irawan, 2024). Secara teknis, *IndoBERT* merupakan model berbasis arsitektur *Transformer* (*BERT*) yang dikembangkan oleh tim peneliti *IndoNLU* (*Indonesian Natural Language Understanding*). Untuk mengakomodasi berbagai tingkat kebutuhan komputasi, arsitektur ini dirilis ke dalam empat varian, yaitu *IndoBERT-Base*, *IndoBERT-Large*, *IndoBERT-liteBase*, dan *IndoBERT-liteLarge*. Dengan varian *IndoBERT-Large* yang dibangun dengan memodifikasi konfigurasi *BERT-Large* (*uncased*) yang dikhususkan untuk tipe bahasa *monolingual*. Arsitektur pada varian ini sangat padat, terdiri dari 24 lapisan tersembunyi (*hidden layers*) dengan dimensi *hidden size* mencapai 1024. Proses komputasi pada lapisan tersebut turut didukung oleh 16 *attention heads* serta jaringan *feed-forward* berdimensi 4096 (Wilie et al., 2020). Dengan spesifikasi arsitektur yang telah dioptimalkan tersebut, *IndoBERT* terbukti memiliki kinerja tinggi untuk mengeksekusi tugas-tugas *NLP*, termasuk analisis sentimen (Asri et al., 2025). Hal ini sejalan dengan temuan (Riyadi et al., 2024) yang mencatat hasil presisi mengesankan dari *IndoBERT*, menegaskan kapasitasnya sebagai instrumen analitis yang sangat tangguh untuk penelitian literatur Indonesia di masa depan.



Gambar 2. 4 Arsitektur Encoder-Only pada IndoBERT

Secara operasional dan arsitektural, seperti yang diperlihatkan pada Gambar 2.4, cara kerja *IndoBERT* sangat berbeda dari model berarsitektur *decoder-only*. *IndoBERT* berfokus secara eksklusif pada tumpukan blok *encoder* dari *Transformer*. Pada setiap blok *encoder* tersebut, terdapat mekanisme *bidirectional self-attention* yang tidak menggunakan penutup (*unmasked*). Hal ini

memungkinkan arsitektur untuk memproses seluruh rangkaian kata dalam suatu kalimat secara serentak dari dua arah sekaligus (bidireksional), baik dari awal ke akhir maupun sebaliknya, sehingga model dapat melihat konteks kalimat secara utuh sejak lapisan pertama. Kemampuan ini terbentuk melalui metode pralatih *Masked Language Modeling* (*MLM*). Saat melakukan inferensi pada tugas klasifikasi sentimen, teks komentar diolah secara paralel melintasi lapisan *attention* dan *feed-forward*. Pada lapisan keluaran paling atas, model tidak menghasilkan token baru secara berurutan, melainkan mengekstrak representasi vektor akhir khusus dari token awal (biasanya token *CLS* atau penanda klasifikasi). Representasi padat ini kemudian disalurkan ke lapisan klasifikasi terpisah (*classification head*) untuk dihitung skor akhirnya menjadi label sentimen positif atau negatif.

2.2.9 *Confusion Matrix*

Confusion Matrix adalah alat evaluasi yang digunakan untuk menilai kinerja model klasifikasi. *Confusion matrix* berupa tabel silang yang mencatat jumlah kemunculan antara dua penilai, klasifikasi yang sebenarnya/aktual dan klasifikasi yang diprediksi (Grandini et al., 2020). Matriks ini membandingkan hasil prediksi model terhadap data sebenarnya dan berisi empat kategori utama yakni *TP* (*True Positive*) model memprediksi positif dan data sebenarnya positif, *TN* (*True Negative*) model memprediksi negatif dan data sebenarnya negatif, *FP* (*False Positive*) model memprediksi positif tetapi data sebenarnya negatif, serta *FN* (*False Negative*) model memprediksi negatif tetapi data sebenarnya positif. Struktur pemetaan metrik klasifikasi ini dijabarkan secara jelas pada Tabel 2.2.

Tabel 2. 2 Confusion Matrix

		Nilai Aktual	
		Negative	***Positive***
Nilai Diprediksi	***Negative***	*True Negative (TN)*	*False Negative (FN)*
	Positive	*False Positive (FP)*	*True Positive (TP)*

2.2.10 *Classification Report*

Classification report adalah alat penting dalam evaluasi performa model klasifikasi, yang memberikan gambaran menyeluruh terhadap hasil prediksi model

pada data uji (Priantoro et al., 2025). Metrik-metrik yang disajikan dalam *classification report* meliputi akurasi, *presisi*, *recall*, dan *F1-score* yang dihitung berdasarkan nilai-nilai pada confusion matrix. Akurasi (*accuracy*) merupakan rasio jumlah prediksi yang benar terhadap jumlah keseluruhan data (Aditama et al., 2024), yang dihitung dengan rumus.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Metrik berikutnya adalah Presisi (*Precision*), yaitu rasio prediksi positif yang benar terhadap seluruh data yang diprediksi positif (Aditama et al., 2024), yang dihitung dengan rumus:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall, yang juga dikenal sebagai *true positive rate*, merupakan rasio prediksi positif yang benar terhadap seluruh data yang sebenarnya positif (Aditama et al., 2024), yang dihitung dengan rumus:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-score digunakan untuk memperoleh ukuran yang seimbang antara presisi dan *recall*. *F1-score* merupakan rata-rata harmonik dari presisi dan *recall* (Aditama et al., 2024), yang dihitung dengan rumus:

$$F1 - score = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

BAB III
METODOLOGI PENELITIAN

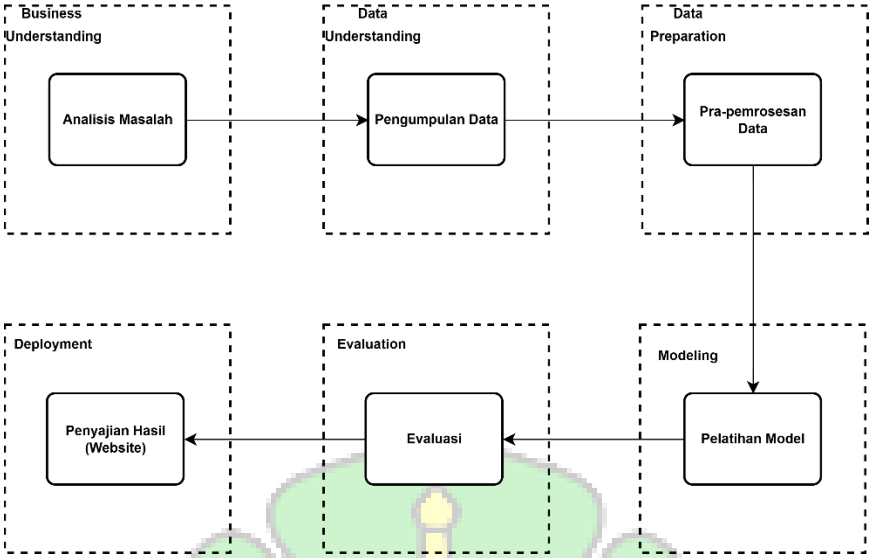
3.1 Jenis Penelitian

Penelitian ini termasuk dalam jenis penelitian eksperimental kuantitatif. Penelitian ini bertujuan untuk menganalisis perbedaan kemampuan tiga model *Natural Language Processing* (*NLP*), yaitu *DeepSeek*, *Gemma*, dan *IndoBERT*, dalam mengklasifikasikan sentimen negatif dan positif pada komentar YouTube topik "Kabur Aja Dulu", serta menentukan model dengan performa keseluruhan terbaik. Evaluasi dilakukan menggunakan *confusion matrix* dan *classification report* yang mencakup *accuracy, precision, recall, F1-score* per kelas. Melalui pendekatan ini diharasspkan dapat dihasilkan tolak ukur kinerja yang objektif untuk masing-masing model.

3.2 Tahapan Penelitian

Prosedur kerja penelitian ini mengadaptasi kerangka kerja *Cross-Industry Standard Process for Data Mining* (*CRISP-DM*). Tahapan dimulai dengan fase Pemahaman Masalah (*Business Understanding*) yakni mendefinisikan permasalahan dan kebutuhan analisis sentimen terkait topik "Kabur Aja Dulu" melalui berbagai Studi Literatur. Dilanjutkan dengan Pemahaman Data (*Data Understanding*) melalui pengumpulan dan eksplorasi data komentar *YouTube*. Kemudian Persiapan Data (*Data Preparation*) yang secara garis besar mencakup pra-pemrosesan, pelabelan sentimen, dan pembagian data. Fase Pemodelan (*Modeling*) berfokus pada implementasi dan pelatihan tiga model yakni *DeepSeek*, *Gemma*, dan *IndoBERT* menggunakan data latih dan validasi. Selanjutnya, pada fase Evaluasi (*Evaluation*), kinerja ketiga model diuji menggunakan data uji pada metrik *confusion matrix* dan *classification report*, hasil evaluasi dari model kemudian dianalisis untuk menilai efektivitas dari model, dan menjadi dasar untuk siklus perbaikan kembali ke fase Pemodelan jika kinerja model dianggap belum optimal. Terakhir, fase *Deployment* yang dalam penelitian ini dilakukan dengan melakukan analisis hasil evaluasi dan penyajian hasil tersebut. Alur dari tahapan-

tahapan penelitian ini dapat dilihat dengan lebih jelas melalui ilustrasi pada Gambar 3.1.



Gambar 3. 1 Tahapan Penelitian

3.2.1 Pemahaman Masalah (*Business Understanding*)

Tahap pertama adalah Pemahaman Masalah, yang berfokus pada pendefinisian fondasi dan tujuan penelitian. Berdasarkan latar belakang, masalah utama yang diidentifikasi adalah adanya topik sosial "Kabur Aja Dulu" yang memicu beragam sentimen di platform *YouTube*. Selain itu, studi literatur secara komprehensif juga dilakukan sebagai bagian dari tahap ini untuk mengidentifikasi keterbatasan penelitian terdahulu dan ruang peningkatan melalui perbandingan model *Large Language Model* yaitu *DeepSeek* dan *Gemma*, dengan model spesifik Bahasa Indonesia, yaitu *IndoBERT*. Penelitian menggunakan pendekatan eksperimental kuantitatif untuk menganalisis perbedaan kemampuan ketiga model dalam mengenali kelas sentimen negatif dan positif berdasarkan metrik per kelas, serta menentukan model dengan performa keseluruhan terbaik berdasarkan metrik evaluasi *confusion matrix* dan *classification report*.

3.2.2 Pemahaman Data (Data Understanding)

Fase kedua, Pemahaman Data, berfokus pada proses pengumpulan dan eksplorasi data awal. Pengumpulan data komentar dilakukan secara otomatis

menggunakan alat *Communalytics* yang diambil dari berbagai video *YouTube* yang relevan dengan kata kunci “Kabur Aja Dulu”. Setelah data mentah terkumpul, dilakukan eksplorasi data awal untuk memahami karakteristik data. Berdasarkan tinjauan terhadap penelitian terdahulu mengenai analisis sentimen media sosial, seperti pada studi pemodelan *Transformer* oleh (Rizkia et al., 2025) yang menggunakan 8.035 data bersih, volume data mentah yang ditargetkan pada penelitian ini ditetapkan minimal 10.000 komentar. Tujuan eksplorasi data adalah untuk memahami karakteristik, dan memastikan relevansi umum komentar dengan topik penelitian sebelum melangkah ke tahap persiapan.

3.2.3 Persiapan Data (*Data Preparation*)

Tahap ketiga, Persiapan Data, merupakan fase intensif yang bertujuan mengubah data mentah yang *noisy* menjadi dataset yang bersih dan siap untuk pemodelan. Seluruh proses pra-pemrosesan pada fase ini dieksekusi secara otomatis menggunakan skrip bahasa pemrograman *Python*. Proses ini diawali dengan pembersihan data (*cleaning*), di mana instruksi kode akan secara otomatis menghapus data duplikat, URL, *mention* (@username), emoji, tanda baca berlebihan, dan karakter non-alfanumerik yang tidak relevan. Selanjutnya, algoritma akan melakukan normalisasi teks, yang mencakup *case folding* (mengubah semua huruf menjadi huruf kecil) dan pemetaan otomatis kata-kata slang atau singkatan yang umum ditemukan di media sosial agar menjadi kata baku. Setelah teks dinormalisasi, proses dilanjutkan dengan penghapusan *stopword* menggunakan *library NLP* untuk menghilangkan kata-kata umum Bahasa Indonesia yang tidak memiliki bobot sentimen, seperti 'di', 'dan', atau 'yang'. Langkah krusial berikutnya adalah pelabelan data (*labeling*), di mana program akan mencocokkan setiap komentar dan memberinya label sentimen (positif dan negatif) secara otomatis dengan memanfaatkan lexicon berbahasa Indonesia. Label otomatis tersebut digunakan sebagai *pseudo-label* atau label referensi untuk proses pelatihan, validasi, dan pengujian, bukan sebagai *ground truth* yang telah dikonfirmasi oleh annotator manusia. Terakhir, dataset yang telah bersih dan berlabel dibagi menjadi tiga bagian, yakni data latih 70% (training set), data validasi 15% (validation set) dan data uji 15% (testing set).

3.2.4 Pemodelan (*Modeling*)

Tahap keempat adalah Pemodelan, di mana eksperimen komputasional untuk melatih dan membangun model klasifikasi sentimen dilakukan sepenuhnya secara otomatis berbasis kode. Penelitian ini akan mengimplementasikan dan mengevaluasi tiga arsitektur model *pre-trained* yang berbeda satu per satu menggunakan dataset yang sama. Model pertama, *IndoBERT* digunakan sebagai baseline yang kuat karena telah dilatih secara ekstensif pada korpus Bahasa Indonesia. Dua model lainnya, *DeepSeek* dan *Gemma*, dipilih sebagai representasi dari *Large Language Model* global generasi baru yang akan diuji kinerjanya pada konteks Bahasa Indonesia informal. Melalui instruksi pemrograman, ketiga model pre-trained ini kemudian akan diadaptasi dan dilatih kembali secara independen untuk melakukan klasifikasi menggunakan data latih dan data validasi yang telah disiapkan. Proses pelatihan terkomputerisasi untuk masing-masing model juga akan melibatkan optimalisasi hyperparameter secara sistematis, seperti *learning rate*, *batch size*, dan jumlah *epoch*, guna mendapatkan performa optimal dari arsitektur tersebut sebelum dilanjutkan ke tahap evaluasi.

Untuk menjaga keadilan interpretasi, ketiga model menggunakan dataset dan pembagian data yang sama. Namun, prosedur pelatihannya tidak sepenuhnya setara: *IndoBERT* dioptimalkan melalui *full fine-tuning* dengan kepala klasifikasi dan *classification loss*, sedangkan *DeepSeek* dan *Gemma* diadaptasi melalui *QLoRA* pada *causal language model* dengan pendekatan *prompt-based generation*. Oleh karena itu, hasil penelitian diposisikan sebagai perbandingan praktis terhadap tiga konfigurasi yang diterapkan, bukan sebagai eksperimen terkontrol yang mengisolasi pengaruh arsitektur, ukuran model, atau metode *fine-tuning* secara tunggal.

3.2.5 Evaluasi (*Evaluation*)

Fase kelima, Evaluasi bertujuan untuk mengukur kinerja objektif dari ketiga model yang telah dilatih secara komputasional. Pengujian ini dieksekusi menggunakan skrip bahasa pemrograman *Python* pada data uji, yaitu data yang belum pernah dilihat oleh model selama proses pelatihan. Kinerja model dihitung menggunakan pustaka evaluasi standar untuk tugas klasifikasi. Metrik yang

digunakan mencakup *confusion matrix* untuk memperlihatkan prediksi benar dan salah pada setiap kelas, *accuracy* untuk mengukur proporsi seluruh prediksi yang tepat, *precision* untuk mengukur ketepatan prediksi suatu kelas, *recall* untuk mengukur kemampuan model menemukan sampel pada kelas tersebut, serta *F1-score* sebagai rata-rata harmonik *precision* dan *recall*. Seluruh metrik dikompilasi dalam *classification report*. Karena distribusi kelas data uji tidak seimbang, penentuan peringkat model tidak didasarkan pada *accuracy* saja. *Macro F1-score* ditetapkan sebagai metrik utama karena memberikan bobot yang sama pada kelas negatif dan positif. *Weighted F1-score* dan *accuracy* digunakan sebagai metrik pendukung, sedangkan *confusion matrix* digunakan untuk menelaah pola kesalahan pada setiap kelas.

Karena label pada data uji berasal dari pelabelan otomatis *InSet*, hasil metrik evaluasi dalam penelitian ini diinterpretasikan sebagai tingkat kesesuaian prediksi model terhadap *pseudo-label InSet*. Nilai tersebut belum dapat disamakan sepenuhnya dengan akurasi terhadap penilaian sentimen manusia karena penelitian ini belum menggunakan anotasi independen oleh annotator manusia

3.2.6 Penyajian Hasil (*Deployment*)

Tahap terakhir adalah Penyajian Hasil (*Deployment*), yang merupakan tahap implementasi agar model yang telah dilatih dan dievaluasi dapat digunakan secara langsung oleh pengguna untuk melakukan klasifikasi sentimen. Sebagai luaran praktis dari penelitian ini dan untuk memenuhi kebutuhan aksesibilitas, hasil akhir penelitian diimplementasikan dalam bentuk aplikasi web interaktif. Berdasarkan perbandingan hasil evaluasi, model dengan performa keseluruhan terbaik dipilih untuk tahap *deployment*, dengan *macro F1-score* sebagai metrik utama serta *weighted F1-score*, *accuracy*, metrik per kelas, dan *confusion matrix* sebagai pendukung. Pemilihan satu model bertujuan menjaga efisiensi komputasi dan waktu respons saat aplikasi digunakan secara *real-time*. Website dibangun menggunakan *Streamlit*, yaitu kerangka kerja *open-source* berbasis *Python* untuk mempublikasikan model *machine learning* sebagai aplikasi web.

BAB IV
HASIL DAN PEMBAHASAN

4.1 Hasil

Bab ini menguraikan hasil dari serangkaian tahapan penelitian yang diimplementasikan berdasarkan kerangka kerja *CRISP-DM*, yang mencakup tahap pemahaman masalah (*business understanding*), pemahaman data (*data understanding*), persiapan data (*data preparation*), pemodelan (*modeling*), evaluasi (*evaluation*), hingga penyajian hasil (*deployment*). Keberhasilan penelitian ini diukur melalui dua lapisan pengujian yang saling melengkapi. Kinerja klasifikasi dari setiap model dievaluasi secara mendalam menggunakan *confusion matrix* dan *classification report*, sementara fungsionalitas serta kelayakan operasional pada aplikasi *Streamlit* divalidasi secara terpisah melalui pengujian *black-box*.

4.1.1 Pemahaman Masalah (*Business Understanding*)

Fenomena "Kabur Aja Dulu" merupakan ekspresi yang ramai di media sosial Indonesia sebagai cerminan keresahan publik terhadap keterbatasan lapangan pekerjaan, biaya pendidikan yang tinggi, dan kondisi ekonomi yang tidak stabil. Tagar ini menjadi simbol kekecewaan terhadap kebijakan pemerintah dan mendorong masyarakat untuk mempertimbangkan mencari peluang hidup yang lebih baik di luar negeri. *YouTube*, sebagai platform video terbesar, menjadi salah satu wadah utama bagi masyarakat untuk menuangkan opini dan sentimen terkait topik ini dalam bentuk komentar. Memahami sentimen publik terhadap isu sosial seperti "Kabur Aja Dulu" menjadi penting bagi sosiolog, jurnalis, dan pemangku kebijakan untuk mengukur dinamika dan persepsi masyarakat secara objektif berbasis data.

Studi literatur yang komprehensif telah dilakukan terhadap sepuluh penelitian terdahulu yang relevan, sebagaimana yang dirangkum pada Tabel 2.1. Penelitian-penelitian tersebut mencakup pendekatan dari model machine learning klasik seperti Bagging Classifier yang mencapai akurasi 95,82%, hingga arsitektur Transformer seperti *BERT* (67%), *RoBERTa* (95%), dan *IndoBERT* (94,6%) pada

berbagai domain dan bahasa. Namun, dari seluruh penelitian yang ditinjau, belum ditemukan studi yang secara spesifik membandingkan kinerja *Large Language Model* (LLM) global generasi baru seperti *DeepSeek* dan *Gemma* dengan model spesifik Bahasa Indonesia *IndoBERT* pada tugas analisis sentimen komentar *YouTube* berbahasa Indonesia yang informal dan penuh slang. Gap inilah yang menjadi pendorong utama penelitian ini.

Berdasarkan identifikasi masalah dan gap penelitian tersebut, dirumuskan dua tujuan utama. Pertama, menganalisis perbedaan kemampuan model *DeepSeek*, *Gemma*, dan *IndoBERT* dalam mengklasifikasikan sentimen negatif dan positif berdasarkan *precision*, *recall*, dan *F1-score* pada setiap kelas. Kedua, menentukan model dengan performa keseluruhan terbaik berdasarkan *accuracy*, *macro F1-score*, *weighted F1-score*, dan *confusion matrix*.

Penelitian ini menggunakan pendekatan eksperimental kuantitatif, di mana tiga arsitektur model yang berbeda diimplementasikan dan dievaluasi secara objektif pada dataset yang sama. Pemilihan ketiga model didasarkan pada representasi dari dua kategori yakni *IndoBERT* sebagai *baseline* yang kuat karena dilatih khusus untuk Bahasa Indonesia, serta *DeepSeek* dan *Gemma* sebagai representasi LLM global generasi baru yang klaim pemahaman konteksnya perlu diuji pada Bahasa Indonesia informal di media sosial. Tahapan *business understanding* ini menjadi landasan yang mengarahkan seluruh proses penelitian, mulai dari pengumpulan data, pemilihan model, hingga metrik evaluasi yang digunakan.

4.1.2 Pemahaman Data (*Data Understanding*)

Dataset yang digunakan dalam penelitian ini bersumber dari hasil ekstraksi komentar publik pada dua video *YouTube* yang relevan dengan topik "Kabur Aja Dulu". Pemilihan kedua video tersebut dilakukan dengan mempertimbangkan kesesuaian isi video dengan topik penelitian serta adanya interaksi pengguna dalam bentuk komentar yang dapat digunakan sebagai sumber data. Video pertama dapat diakses melalui tautan *https://www.youtube.com/watch?v=AaMMc0mhMKw* dan video kedua pada tautan *https://www.youtube.com/watch?v=_xxvRK6_AI8*. Proses pengumpulan data dilakukan secara otomatis menggunakan alat *Communalytics* yang mengkhususkan diri pada ekstraksi komentar *YouTube*. Dari proses

pengumpulan awal, diperoleh data mentah (*raw data*) sebanyak 19.189 komentar yang berasal dari gabungan kedua video tersebut. Jumlah ini melampaui target minimal 10.000 komentar mentah yang ditetapkan pada Bab III, setelah tahap persiapan data, dataset untuk pemodelan kemudian disampling menjadi 10.000 komentar.

Eksplorasi data awal dilakukan untuk memahami karakteristik umum dari data yang terkumpul, meliputi struktur kolom, tipe data, keberadaan nilai kosong, serta distribusi sentimen awal. Hasil eksplorasi menunjukkan bahwa data mentah didominasi secara signifikan oleh komentar dengan sentimen negatif, sementara komentar positif hanya mencakup sebagian kecil dari keseluruhan dataset. Ketimpangan ini menjadi perhatian utama karena dapat menyebabkan model bias terhadap kelas mayoritas jika tidak ditangani dengan baik pada tahap pra-pemrosesan.

4.1.3 Persiapan Data (*Data Preparation*)

Tahap persiapan data merupakan fase intensif yang bertujuan mengubah data mentah menjadi dataset yang bersih dan siap untuk pemodelan. Seluruh proses pra-pemrosesan pada fase ini dieksekusi secara otomatis menggunakan skrip bahasa pemrograman *Python* pada platform *Google Colab*. Data dari dua file *Excel* yang telah diunggah dan digabungkan menjadi 19.189 komentar melalui serangkaian tahapan yang sistematis. Tabel 4.1 menyajikan contoh sampel dataset setelah seluruh tahapan pra-pemrosesan.

Tabel 4. 1 Sampel Dataset Hasil Pra-pemrosesan

Teks Asli	Teks Bersih	Sentimen	Skor Lexicon	Split
Ilmuwan doktor sarjana insiyur.pada kabur.itu banyak faktanya.yg berkuasa indo ya koruptor	ilmuwan doktor sarjana insiyur pada kabur itu banyak faktanya yang berkuasa indo ya koruptor	*negative*	-8.0	train

alhammdllah moga makin bnyak yg kabur,lumayan ngurangin beban pemerintah	alhammdllah moga makin bnyak yang kabur lumayan ngurangin beban pemerintah	*negative*	-14.0	train
mestinya semangat '45 DIBANGKITKAN KEMBALI .... mengusir asing	mestinya semangat 45 dibangkitkan kembali mengusir asing	*negative*	-12.0	test
Jepang kaga ada nih?	jepang kaga ada nih	*positive*	4.0	valid

Berikut adalah alur pra-pemrosesan yang dilakukan secara berurutan:

1) Pembersihan Data dan Deduplikasi

   Tahap pertama adalah pembersihan data dan deduplikasi. Proses ini diawali dengan menghapus baris yang tidak memiliki teks (kosong), menghapus duplikasi berdasarkan ID komentar, serta menghapus duplikasi teks yang benar-benar identik secara global. Selain itu, teks yang hasil pembersihannya menjadi kosong juga dihapus dari dataset. Hasil dari tahap ini mengurangi data dari 19.189 menjadi 18.808 komentar unik, dengan rincian 14 data hilang karena teks bersih kosong dan 367 data hilang karena duplikasi teks

2) Normalisasi Teks

   Tahap kedua adalah normalisasi teks (*text cleaning*) yang komprehensif. Proses ini memperbaiki anomali format teks menggunakan *ftfy* dan *html.unescape*, serta menormalisasi karakter *Unicode*. Elemen yang tidak relevan untuk analisis sentimen seperti tautan (URL), *mention* (simbol @) dihapus, sementara hashtag dipertahankan konten katanya. Emoji dikonversi menjadi representasi teks (*demojize*). Seluruh teks kemudian diubah menjadi huruf kecil (*case folding*), tanda baca dihapus, dan karakter huruf yang berulang secara berlebihan direduksi. Selanjutnya, normalisasi bahasa gaul (*slang normalization*) dilakukan dengan mengonversi kata-kata

tidak baku seperti "yg" menjadi "yang", "tdk" menjadi "tidak", "gw" menjadi "saya", dan puluhan kata slang lainnya ke bentuk baku menggunakan kamus yang telah ditentukan.

3) *Stemming*

Tahap ketiga adalah *stemming* menggunakan pustaka *Sastrawi*. Proses ini mengubah setiap kata ke dalam bentuk dasarnya (*root word*) untuk menyederhanakan variasi morfologis kata. Sebagai contoh, kata "selengkapnya" menjadi "lengkap", "kerusuhan" menjadi "rusuh", dan "berlalu" menjadi "lalu". Stemming bertujuan untuk mereduksi dimensi fitur dengan menyatukan varian kata yang memiliki makna dasar yang sama.

4) Pelabelan Sentimen Berbasis *Lexicon* (*Pseudo-labeling*)

Tahap keempat adalah pelabelan sentimen otomatis (*lexicon-based labeling*). Proses ini menggunakan kamus leksikon *InSet* (*Indonesian Sentiment Lexicon*) yang menurut (Koto & Rahmaningtyas, 2017) memuat 3.609 entri positif dan 6.609 entri negatif, sehingga totalnya 10.218 entri dengan bobot polaritas antara -5 dan +5. Algoritma pelabelan menghitung skor sentimen dengan mempertimbangkan beberapa faktor, yaitu deteksi kata negasi (seperti "tidak", "bukan") dalam jendela 3 kata sebelumnya yang dapat membalikkan skor sentimen, kata penguat (intensifier seperti "sangat", "banget") yang menggandakan skor 1,5 kali, serta kata pelemah (diminisher seperti "agak", "sedikit") yang mengurangi skor menjadi 0,75 kali. Hasil pelabelan menunjukkan bahwa dari 18.808 komentar, sebanyak 14.848 (78,95%) berlabel negatif, 2.868 (15,25%) berlabel positif, dan 1.092 (5,81%) berlabel netral. Label yang dihasilkan *InSet* selanjutnya digunakan sebagai *pseudo-label* referensi pada tahap pemodelan dan evaluasi. Penelitian ini belum mencakup validasi manual oleh annotator independen, sehingga *pseudo-label* tersebut berpotensi mengandung noise, terutama pada komentar yang bersifat sarkastik, implisit, ambigu, atau sangat bergantung pada konteks.

5) Penyaringan dan Pemisahan Data

Tahap kelima adalah penyaringan dan pemisahan data. Komentar yang terindikasi bersentimen netral dihapus dari dataset, sehingga hanya menyisakan data biner (positif dan negatif) sebanyak 17.716 komentar dengan distribusi 83,81% negatif dan 16,19% positif. Untuk memperbaiki ketidakseimbangan kelas yang masih cukup tinggi, diterapkan strategi sampling dengan mengambil seluruh data positif sebanyak 2.868 komentar dan menambahkan data negatif sebanyak 7.132 komentar, sehingga menghasilkan dataset final berjumlah 10.000 data dengan distribusi 71,32% negatif dan 28,68% positif. Dataset ini kemudian dibagi ke dalam tiga bagian dengan rasio 70:15:15 menggunakan metode stratified split untuk mempertahankan proporsi kelas di setiap subset. Pada data latih terdapat 4.992 data negatif (71,31%) dan 2.008 data positif (28,69%), sedangkan pada data validasi dan data uji masing-masing terdiri dari 1.070 data negatif (71,33%) dan 430 data positif (28,67%).

Dataset yang tidak seimbang ini menjadi pertimbangan penting dalam pemilihan metrik evaluasi model. Jika hanya menggunakan metrik akurasi, model dapat terlihat memiliki performa yang baik karena mayoritas data berasal dari kelas negatif, namun model tersebut belum tentu mampu mengenali kelas positif dengan baik. Oleh karena itu, evaluasi dalam penelitian ini tidak hanya menggunakan akurasi, tetapi juga menggunakan metrik *weighted F1-score*, *macro F1-score*, *precision*, *recall*, *F1-score* per kelas, serta *confusion matrix* untuk memberikan gambaran performa yang lebih komprehensif.

4.1.4 Pemodelan (*Modeling*)

Tahap pemodelan merupakan fase implementasi dan pelatihan tiga arsitektur model pre-trained untuk tugas klasifikasi sentimen biner (positif dan negatif). Ketiga model dilatih menggunakan dataset yang sama dengan pembagian 70% data latih, 15% data validasi, dan 15% data uji. Setiap model memiliki pendekatan *fine-tuning* yang berbeda yakni *IndoBERT* menggunakan *full fine-tuning* sebagai model klasifikasi sekuens, sedangkan DeepSeek dan *Gemma* menggunakan pendekatan *QLoRA* dengan *prompt-based generation* pada *causal language model*.

Penggunaan dataset dan *split* yang sama memberikan dasar perbandingan yang konsisten, tetapi beberapa faktor eksperimen tetap berbeda, meliputi *objective* atau *loss*, jumlah parameter yang diperbarui, kuantisasi, panjang maksimum *token*, *batch size*, *learning rate*, serta mekanisme keluaran model. Dengan demikian, perbedaan performa dapat dipengaruhi oleh gabungan faktor tersebut. Interpretasi hasil pada penelitian ini dibatasi pada konfigurasi eksperimen yang digunakan dan tidak dimaksudkan sebagai pembuktian bahwa satu keluarga arsitektur selalu lebih unggul dalam seluruh kondisi.

1) *IndoBERT*

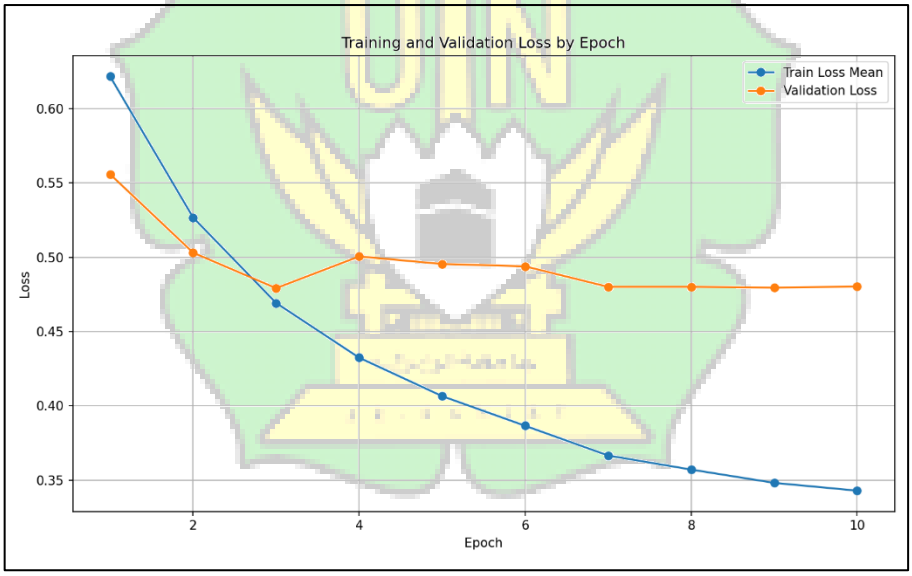
   Model *IndoBERT* yang digunakan adalah varian i*ndobenchmark/indobert-base-p1*, yaitu arsitektur *Transformer* berbasis *BERT* yang dilatih khusus untuk bahasa Indonesia. *IndoBERT* digunakan sebagai baseline dalam penelitian ini karena telah terbukti memiliki kinerja tinggi pada berbagai tugas *NLP* Bahasa Indonesia. Rincian pengaturan parameter untuk model ini disajikan pada Tabel 4.2.

Tabel 4. 2 Konfigurasi IndoBERT

Parameter	**Nilai**
Model dasar	*indobenchmark/indobert-base-p1*
Arsitektur	*BERT* (*sequence classification*)
Hidden layers	12
Hidden size	768
Attention heads	12
Intermediate size	3072
Dropout	0.3
Maksimal panjang token	64
Pendekatan training	*Full fine-tuning*
Batch size (*train/eval*)	32 / 32
Learning rate	7e-6
Learning rate scheduler	*Cosine*
Warmup ratio	0.1
Weight decay	0.2

Label smoothing	0.1
Max grad norm	1.0
Epoch	10
Optimizer	*AdamW* (*FP16*)
Metrik pemilihan model	f1_weighted

Proses fine-tuning *IndoBERT* dilakukan dengan pendekatan *full fine-tuning*, di mana seluruh parameter model diperbarui selama pelatihan. Model dikonfigurasi ulang dengan num_labels = 2 dan pemetaan label2id = {"negative": 0, "positive": 1}. *Tokenizer* yang digunakan adalah *tokenizer* bawaan *IndoBERT* dengan padding ke panjang maksimal dan *truncation* pada 64 token. Pelatihan menggunakan optimizer *AdamW* dengan *mixed precision FP16* untuk mempercepat komputasi. Best model dipilih berdasarkan metrik f1_weighted tertinggi pada data validasi.



Gambar 4. 1 Kurva Training dan Validation Loss IndoBERT per Epoch

Seperti yang ditunjukkan pada Gambar 4.1, hasil *training convergence IndoBERT* menunjukkan pola pembelajaran yang stabil. *Training loss* menurun secara konsisten dari 0,62 pada *epoch* 1 menjadi 0,34 pada *epoch* 10. *Validation loss* juga menurun dari 0,56 menjadi 0,48 dan mulai stabil pada *epoch* 7 hingga 10.

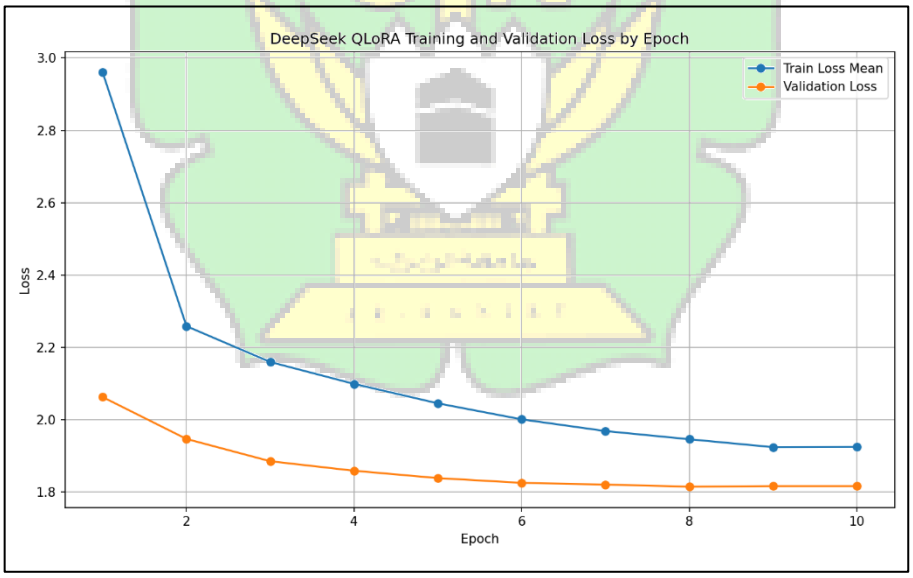
2) *Deepseek*

Model *DeepSeek* yang digunakan adalah *deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B*, yaitu *causal language model dense* dengan arsitektur *decoder-only* berbasis *Qwen2* dan skala sekitar 1,5 miliar parameter. Model ini bukan *DeepSeek-R1* penuh yang menggunakan arsitektur *Mixture-of-Experts*. Konfigurasi inti *checkpoint* terdiri dari 28 lapisan *Transformer*, *hidden size* 1.536, 12 *attention heads*, dan 2 *key-value heads*. Untuk menyesuaikan proses *fine-tuning* dengan keterbatasan memori komputasi, model diadaptasi menggunakan pendekatan *QLoRA* (*Quantized Low-Rank Adaptation*). Konfigurasi lengkap model ini dapat dilihat pada Tabel 4.3.

Tabel 4. 3 Konfigurasi DeepSeek

Parameter	Nilai
Model dasar	*deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B*
Arsitektur	*Causal LM* (*decoder-only Transformer* berbasis *Qwen2*, *dense*)
Pendekatan training	*QLoRA* (*4-bit*)
Kuantisasi	*NF4*, *double quant*, *bfloat16*
LoRA rank (r)	8
LoRA alpha	16
LoRA dropout	0.05
Target modul *LoRA*	q_proj, k_proj, v_proj, o_proj
Batch size per device	1
Gradient accumulation	16 (efektif 16)
Learning rate	2e-4
Learning rate scheduler	*Cosine*
Warmup ratio	0.03
Weight decay	0.01
Maksimal panjang token	256
Epoch	10
Optimizer	paged_adamw_8bit (BF16)
Pendekatan klasifikasi	*Prompt-based generation*

Proses pelatihan *DeepSeek* menggunakan pendekatan *QLoRA* dengan kuantisasi 4-bit *NF4* dan *double quantization* untuk menghemat memori. Model dasar dimuat dalam 4-bit dengan *compute dtype bfloat16*. *LoRA* adapter dengan *rank* 8 ditambahkan pada modul *query*, *key*, *value*, dan *output projection*. Hanya parameter *adapter* yang diperbarui selama pelatihan, sementara parameter model dasar tetap dalam kondisi terkuantisasi. Klasifikasi sentimen dilakukan melalui pendekatan *prompt-based generation*, di mana teks komentar dimasukkan ke dalam *template prompt*: "*Analyze the sentiment of the Indonesian sentence in [brackets]. Return exactly one label: 1 for positive, 0 for negative. [text]* =". Model kemudian menghasilkan teks prediksi yang di-parse menjadi label 0 atau 1. *Batch size* 1 per *device* dikompensasi dengan gradient accumulation sebanyak 16 langkah sehingga menghasilkan *effective batch size* 16. *Optimizer* yang digunakan adalah paged_adamw_8bit untuk mengelola memori secara efisien.



Gambar 4. 2 Kurva Training dan Validation Loss DeepSeek per Epoch

Berdasarkan Gambar 4.2, hasil *training convergence DeepSeek* menunjukkan bahwa *training loss* menurun dari 2,96 pada *epoch* 1 menjadi 1,92 pada *epoch* 10. *Validation loss* menurun dari 2,06 pada *epoch* 1 menjadi 1,82 pada *epoch* 8 dan mulai stabil hingga *epoch* 10. *Validation loss* yang stabil menunjukkan bahwa model tidak mengalami *overfitting*

yang signifikan, meskipun performa klasifikasi aktualnya baru dapat diketahui melalui evaluasi pada data uji.

3) Gemma

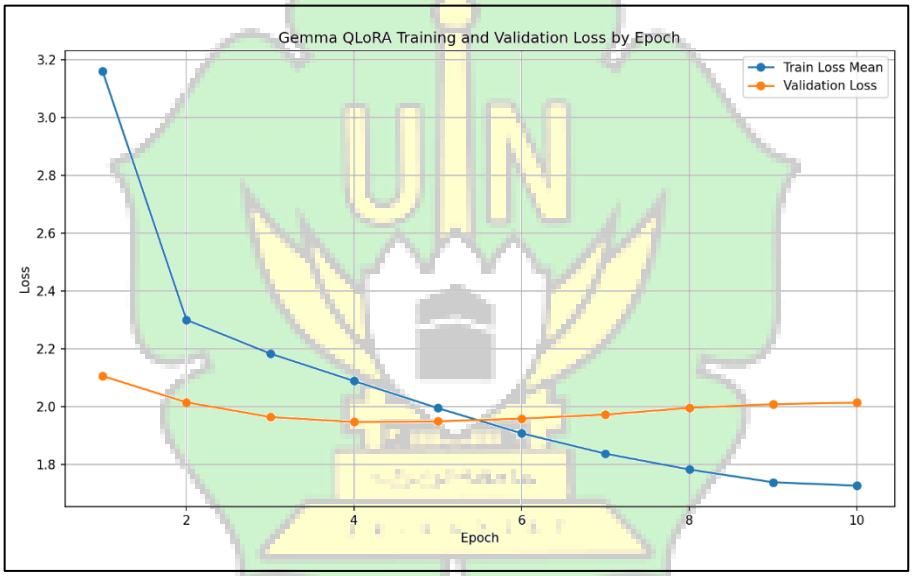
Model *Gemma* yang digunakan adalah *google/gemma-1.1-2b-it*, yaitu versi *instruction-tuned* dari *Gemma 1.1* dengan sekitar 2 miliar parameter. Model ini merupakan *text-to-text causal language model* dengan arsitektur *decoder-only Transformer* yang bersifat *dense*. Konfigurasi dasarnya terdiri dari 18 lapisan *Transformer*, *hidden size* 2.048, 8 *query attention heads*, dan 1 *key-value head* melalui mekanisme *Multi-Query Attention*. Sama seperti *DeepSeek*, *Gemma* diadaptasi menggunakan pendekatan *QLoRA* untuk menyesuaikan kebutuhan memori selama proses pelatihan. Pengaturan parameter secara lengkap disajikan pada Tabel 4.4.

Tabel 4. 4 Konfigurasi Gemma

Parameter	Nilai
Model dasar	*google/gemma-1.1-2b-it*
Arsitektur	*Causal LM* (*decoder-only Transformer Gemma*, *dense*)
Pendekatan training	*QLoRA* (4-bit)
Kuantisasi	*NF4*, *double quant*, *bfloat16*
LoRA rank (r)	8
LoRA alpha	16
LoRA dropout	0.05
Target modul *LoRA*	q_proj, k_proj, v_proj, o_proj
Batch size per device	1
Gradient accumulation	16 (efektif 16)
Learning rate	2e-4
Learning rate scheduler	*Cosine*
Warmup ratio	0.03
Weight decay	0.01
Maksimal panjang token	256
Epoch	10

Optimizer	paged_adamw_8bit (BF16)
Pendekatan klasifikasi	*Prompt-based generation*

Proses pelatihan *Gemma* menggunakan pendekatan *QLoRA* yang sama dengan *DeepSeek*, yaitu kuantisasi 4-bit *NF4*, *double quantization*, dan *LoRA adapter rank* 8 pada modul *attention projection*. Klasifikasi dilakukan melalui *prompt-based generation* dengan *template prompt* yang sama. Perbedaan utama terletak pada arsitektur model dasar dan skala parameter dengan *Gemma* memiliki sekitar 2 miliar parameter, sedangkan *DeepSeek-R1-Distill-Qwen* memiliki sekitar 1,5 miliar parameter. Kedua model tetap dikompresi dalam format 4-*bit* dan dilatih dengan *effective batch size* yang sama agar sesuai dengan keterbatasan sumber daya komputasi.



Gambar 4. 3 Kurva Training dan Validation Loss Gemma per Epoch

Seperti yang diilustrasikan pada Gambar 4.3, hasil *training convergence Gemma* menunjukkan pola yang berbeda dibandingkan kedua model lainnya. Training loss terus menurun secara konsisten dari 3,16 pada *epoch* 1 menjadi 1,73 pada *epoch* 10. Namun, *validation loss* menunjukkan pola yang mengkhawatirkan yakni menurun dari 2,11 pada *epoch* 1 menjadi 1,95 pada *epoch* 4, namun kemudian mulai meningkat kembali menjadi 2,01 pada *epoch* 10. Peningkatan *validation loss* setelah *epoch* 4 sementara *training loss* terus menurun merupakan indikasi klasik *overfitting*, di mana

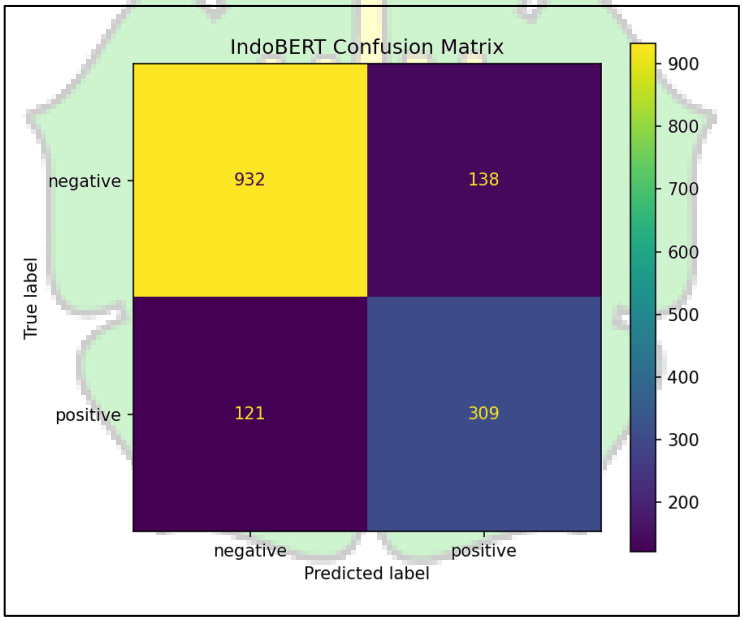
model mulai menghafal pola pada data latih dan kehilangan kemampuan generalisasi terhadap data baru.

4.1.5 Evaluasi (*Evaluation*)

Tahap evaluasi bertujuan untuk mengukur dan membandingkan kinerja ketiga model yang telah dilatih menggunakan data uji yang terdiri dari 1.500 komentar (1.070 negatif dan 430 positif). Model dievaluasi menggunakan metrik *Confusion Matrix* dan *Classification Report* yang mencakup akurasi, *presisi*, *recall*, dan *F1-score*.

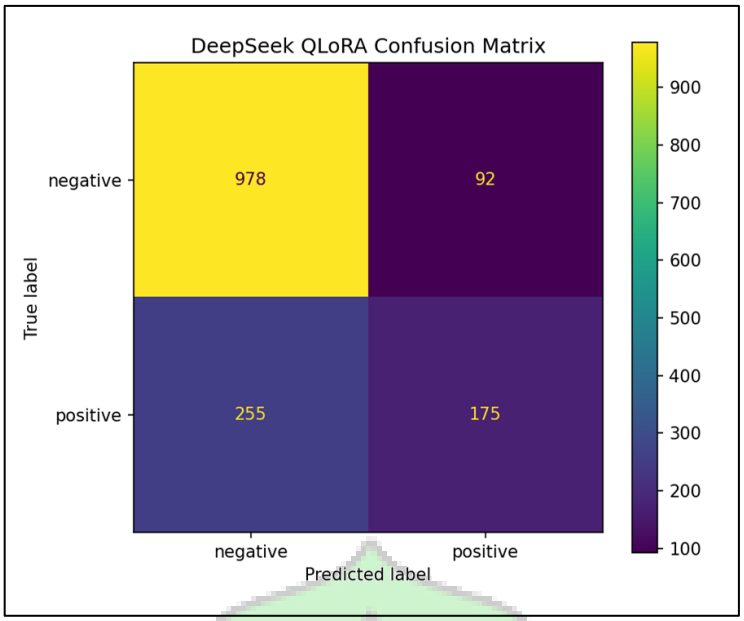
1) *Confusion Matrix*

   Confusion Matrix menyajikan jumlah prediksi benar dan salah pada setiap kelas untuk masing-masing model.



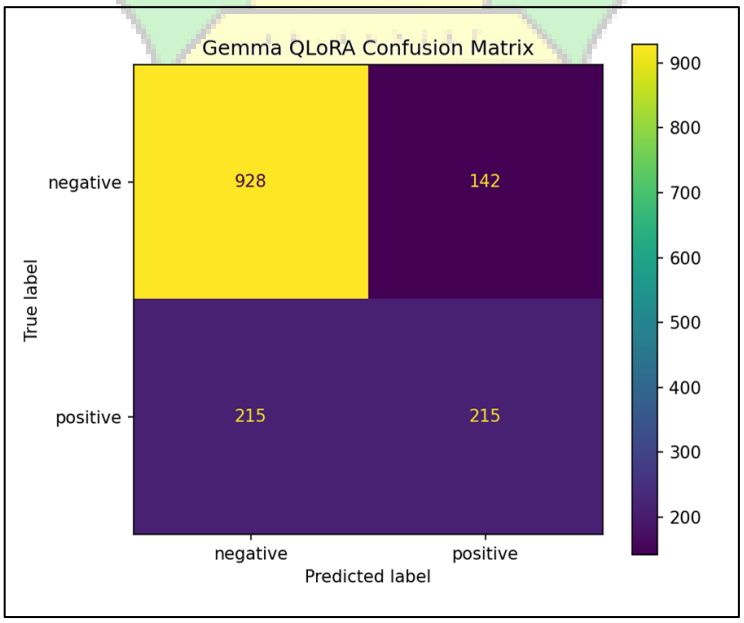
Gambar 4. 4 Confusion Matrix IndoBERT

Berdasarkan Gambar 4.4, *IndoBERT* berhasil mengklasifikasikan 932 sampel negatif dan 309 sampel positif dengan benar. Model ini menghasilkan 138 *false positive* dan 121 *false negative*. Jumlah *false negative* yang paling rendah di antara ketiga model menunjukkan bahwa IndoBERT memiliki kemampuan terbaik dalam mengenali sentimen positif.



Gambar 4. 5 Confusion Matrix DeepSeek

Seperti yang terlihat pada Gambar 4.5, *DeepSeek* berhasil mengklasifikasikan 978 sampel negatif dan 175 sampel positif dengan benar. Model ini menghasilkan 92 *false positive* (paling rendah di antara ketiga model) namun 255 *false negative* (paling tinggi di antara ketiga model). Hal ini menunjukkan bahwa *DeepSeek* sangat cenderung memprediksi data sebagai negatif, sehingga kuat dalam mengenali kelas negatif tetapi lemah dalam mendeteksi sentimen positif.



Gambar 4. 6 Confusion Matrix Gemma

Merujuk pada Gambar 4.6, *Gemma* berhasil mengklasifikasikan 928 sampel negatif dan 215 sampel positif dengan benar. Model ini menghasilkan 142 *false positive* dan 215 *false negative*. Performa *Gemma* berada di antara *IndoBERT* dan *DeepSeek*, dengan kemampuan yang cukup baik dalam mengenali kelas positif meskipun masih kalah dibandingkan *IndoBERT*.

2) *Classification Report*

Classification Report menyajikan metrik evaluasi secara terperinci untuk setiap kelas pada masing-masing model.

Tabel 4. 5 Classification Report IndoBERT

Metrik	***Precision***	***Recall***	***F1-Score***
Negatif	0,8851	0,8710	0,8780
Positif	0,6913	0,7186	0,7047
Accuracy			0,8273
Macro Avg	0,7882	0,7948	0,7913
Weighted Avg	0,8295	0,8273	0,8283

Berdasarkan data pada Tabel 4.5, *IndoBERT* memperoleh akurasi tertinggi sebesar 0,8273 dengan *weighted F1-score* 0,8283 dan *macro F1-score* 0,7913. Performa pada kelas positif menunjukkan presisi 0,6913 dan *recall* 0,7186, yang merupakan nilai tertinggi di antara ketiga model. Hal ini mengindikasikan bahwa *IndoBERT* memiliki keseimbangan terbaik dalam mengklasifikasikan sentimen negatif maupun positif.

Tabel 4. 6 Classification Report DeepSeek

Metrik	***Precision***	***Recall***	***F1-Score***
Negatif	0,7932	0,9140	0,8493
Positif	0,6554	0,4070	0,5022
Accuracy			0,7687
Macro Avg	0,7243	0,6605	0,6757
Weighted Avg	0,7537	0,7687	0,7498

Seperti yang ditunjukkan pada Tabel 4.6, *DeepSeek* memiliki keunggulan pada *recall* kelas negatif tertinggi (0,9140) dan *false positive* paling rendah,

yang menunjukkan model ini sangat konservatif dalam memprediksi sentimen positif. Namun, *recall* positif yang sangat rendah (0,4070) mengindikasikan bahwa *DeepSeek* sering gagal mengenali komentar positif dan cenderung mengklasifikasikan hampir semua data sebagai negatif. Akurasi yang diperoleh sebesar 0,7687.

Tabel 4. 7 Classification Report Gemma

Metrik	*Precision*	*Recall*	*F1-Score*
Negatif	0,8119	0,8673	0,8387
Positif	0,6022	0,5000	0,5464
Accuracy			0,7620
Macro Avg	0,7071	0,6836	0,6925
Weighted Avg	0,7518	0,7620	0,7549

Merujuk pada Tabel 4.7 dengan *macro F1-score* sebagai metrik utama, Gemma menempati peringkat kedua dan *DeepSeek* peringkat ketiga. *Gemma* memiliki *recall* positif 0,5000 dan *macro F1-score* 0,6925, lebih tinggi daripada *DeepSeek* yang memperoleh *recall* positif 0,4070 dan *macro F1-score* 0,6757. Namun, apabila peringkat hanya didasarkan pada *accuracy*, *DeepSeek* berada di atas *Gemma* karena memperoleh 0,7687 dibandingkan 0,7620. Perbedaan urutan ini menunjukkan bahwa *accuracy* saja belum cukup untuk menilai model pada data yang tidak seimbang. *Gemma* memberikan keseimbangan antarkelas yang lebih baik, sedangkan *DeepSeek* lebih kuat dalam mengenali kelas negatif.

4.1.6 Penyajian Hasil (*Deployment*)

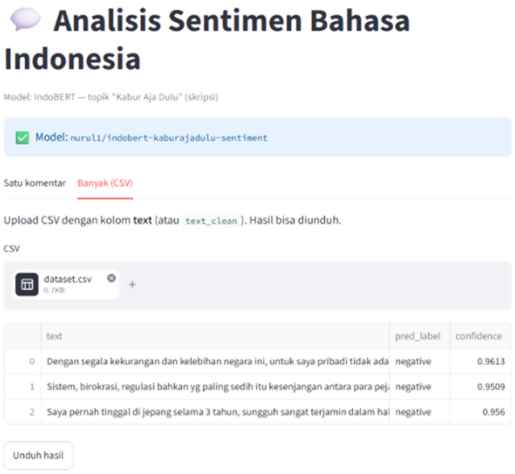
Menindaklanjuti hasil evaluasi yang menetapkan *IndoBERT* sebagai model dengan performa terbaik, tahap penyajian hasil direalisasikan melalui pembuatan aplikasi web interaktif. Pengembangan aplikasi ini dibangun menggunakan kerangka kerja *Streamlit*. Aplikasi ini dirancang untuk memberikan kemudahan bagi pengguna akhir dalam melakukan analisis sentimen komentar *YouTube* pada topik Kabur Aja Dulu secara praktis.

Aplikasi analisis sentimen ini memiliki dua fungsionalitas utama yang dapat diakses melalui tab menu, yaitu pengujian teks tunggal dan pengujian massal. Pada

fitur pengujian teks tunggal atau fitur Satu komentar, pengguna dapat memasukkan satu kalimat ke dalam kolom masukan teks yang disediakan. Setelah pengguna menekan tombol analisis, sistem di latar belakang secara otomatis menjalankan alur pra-pemrosesan teks yang identik dengan fase pelatihan. Teks tersebut kemudian diproses ke dalam model *pre-trained IndoBERT* untuk mengeksekusi proses inferensi. Hasil prediksinya ditampilkan seketika dalam bentuk klasifikasi kelas sentimen beserta persentase tingkat keyakinan probabilitas model, seperti yang ditunjukkan pada Gambar 4.7.

Gambar 4. 7 Tampilan Antarmuka Analisis Sentimen untuk Satu Komentar

Pengguna dapat beralih ke tab Banyak (*CSV*) dan mengunggah dataset berekstensi *CSV* yang memuat teks komentar. Aplikasi akan mengevaluasi keseluruhan baris teks secara otomatis dan menyajikan hasil prediksinya dalam bentuk tabel terstruktur. Tabel keluaran ini mendeskripsikan teks masukan, label prediksi sentimen pada kolom *pred_label*, beserta nilai probabilitasnya pada kolom *confidence*. Untuk kebutuhan perekaman data lanjutan, hasil klasifikasi massal ini juga dapat disimpan dan diunduh oleh pengguna, sebagaimana diilustrasikan pada Gambar 4.8.



Gambar 4. 8 Tampilan Antarmuka Analisis Sentimen Massal Berbasis CSV

Selain verifikasi tampilan antarmuka, aplikasi diuji secara fungsional menggunakan metode *black-box* untuk memastikan setiap fitur memberikan keluaran yang sesuai dengan skenario masukan. Pengujian mencakup masukan teks tunggal, validasi masukan kosong, serta pemrosesan berkas *CSV*. Matriks perancangan dan hasil pengujian aplikasi tersebut disajikan secara lengkap pada Tabel 4.8.

Tabel 4. 8 Pengujian Black-Box Aplikasi Streamlit

Skenario	Input	Hasil yang Diharapkan	Hasil Aktual	Status
Analisis teks positif	"Saya senang melihat banyak anak muda memperoleh peluang yang lebih baik."	Sistem menampilkan label positif dan nilai *confidence* tanpa *error*.	Label positif tampil dengan nilai *confidence* 84,5%.	Lulus
Analisis teks negatif	"Kondisi pekerjaan sekarang sangat buruk dan membuat	Sistem menampilkan label negatif dan nilai *confidence* tanpa *error*.	Label negatif tampil dengan nilai *confidence* 92,1%.	Lulus

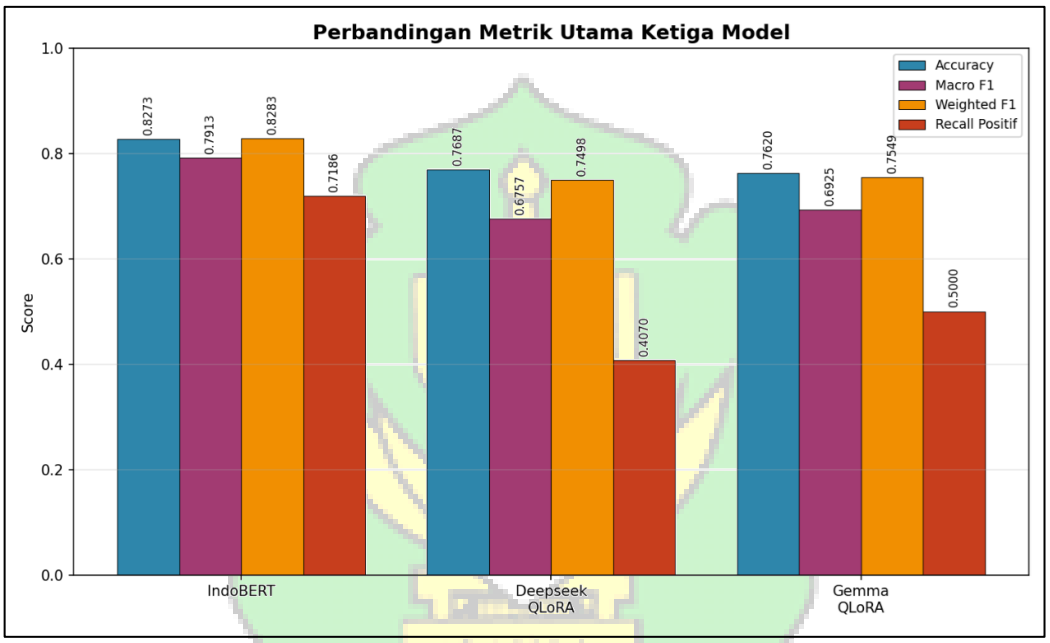
	masyarakat kecewa.”			
Kolom teks kosong	Kolom komentar tidak diisi, kemudian tombol Analisis ditekan.	Sistem menolak proses inferensi dan menampilkan pesan agar pengguna mengisi teks.	Pesan peringatan “Text kosong” tampil.	Lulus
Unggah *CSV* valid	Berkas *comments.csv* berisi kolom text dengan tiga baris komentar.	Sistem membaca seluruh baris, menampilkan *pred_label* dan *confidence*, serta menyediakan hasil unduhan.	Tiga baris berhasil diproses dan tabel hasil dapat diunduh	Lulus
CSV dengan nama kolom salah	Berkas *invalid.csv* tanpa kolom text.	Sistem menolak pemrosesan dan menampilkan informasi mengenai kolom yang diwajibkan.	Pesan error “*CSV* harus punya kolom text atau text_clean.” tampil.	Lulus
Format berkas tidak didukung	Pengguna mencoba mengunggah berkas *comments.xlsx*.	Sistem membatasi unggahan pada format *CSV*.	Berkas *XLSX* tidak diterima oleh komponen unggah.	Lulus

Deployment ini membuktikan bahwa arsitektur *IndoBERT* tidak sekadar unggul dalam metrik performa komputasi secara teoritis, tetapi juga dapat diintegrasikan secara mulus ke dalam sistem perangkat lunak yang nyata. Implementasi web ini berhasil memenuhi tujuan pengembangan dengan menyediakan fungsionalitas operasional untuk mengidentifikasi opini publik terhadap isu sosial secara interaktif dan efisien.

4.2 Pembahasan

Tabel 4. 9 Rekapitulasi Hasil Evaluasi

Model	*Precision*		*Recall*		*F1-Score*		*Accuracy*
	Positif	**Negatif**	**Positif**	**Negatif**	**Positif**	**Negatif**	
IndoBERT	0,6913	0,8851	0,7186	0,8710	0,7047	0,8780	0,8273
DeepSeek	0,6554	0,7932	0,4070	0,9140	0,5022	0,8493	0,7687
Gemma	0,6022	0,8119	0,5000	0,8673	0,5464	0,8387	0,7620



Gambar 4. 9 Perbandingan Metrik Utama

Berdasarkan hasil evaluasi yang dirangkum pada Tabel 4.9 dan Gambar 4.9, peringkat performa keseluruhan ditentukan terutama menggunakan *macro F1-score* karena data uji memiliki distribusi kelas yang tidak seimbang. Dengan kriteria tersebut, urutan model adalah *IndoBERT* pada peringkat pertama dengan *macro F1-score* 0,7913, *Gemma* pada peringkat kedua dengan 0,6925, dan *DeepSeek* pada peringkat ketiga dengan 0,6757. *IndoBERT* juga memperoleh *accuracy* tertinggi sebesar 0,8273 dan *weighted F1-score* tertinggi sebesar 0,8283. Keunggulan paling signifikan terlihat pada kemampuan *IndoBERT* mendeteksi kelas positif, dengan recall 0,7186 dan F1-score 0,7047. Untuk dua model lainnya, urutan berdasarkan *accuracy* berbeda yakni *DeepSeek* memperoleh 0,7687, sedikit lebih tinggi daripada *Gemma* sebesar 0,7620.

Keunggulan *IndoBERT* pada hasil eksperimen ini dapat berkaitan dengan kesesuaian arsitektur dan tujuan optimasi terhadap tugas klasifikasi. *IndoBERT* merupakan model *encoder-based* yang menggunakan kepala klasifikasi dan *classification loss*, sehingga pembaruan parameternya diarahkan secara langsung pada pemisahan kelas sentimen. Sebaliknya, *DeepSeek* dan *Gemma* merupakan *causal language model* yang diadaptasi menggunakan *QLoRA* dan menghasilkan label melalui *prompt-based generation*. Meskipun demikian, perbedaan hasil tidak dapat diatribusikan hanya kepada arsitektur *encoder* dan *decoder* karena ketiga konfigurasi juga berbeda dalam metode *fine-tuning*, *objective*, kuantisasi, panjang *token*, *batch size*, *learning rate*, dan mekanisme keluaran. Penjelasan ini karena itu bersifat interpretatif, bukan pembuktian kausal terhadap satu faktor tunggal.

DeepSeek menunjukkan kelemahan paling mencolok pada *recall* positif yang rendah, yaitu 0,4070, sehingga model hanya mengenali 40,70% dari seluruh sampel positif. *DeepSeek* cenderung konservatif dalam memprediksi sentimen positif, terlihat dari *false positive* paling rendah sebanyak 92, tetapi *false negative* paling tinggi sebanyak 255. *Gemma* menunjukkan keseimbangan yang lebih baik dengan *recall* positif 0,5000 dan *macro F1-score* 0,6925, meskipun tetap berada di bawah *IndoBERT*. Kurva pelatihan *Gemma* memperlihatkan *validation loss* yang meningkat setelah *epoch* 4 ketika training loss terus menurun. Pola tersebut konsisten dengan indikasi *overfitting* dan dapat berkaitan dengan penurunan generalisasi, tetapi data yang tersedia belum cukup untuk menyimpulkan hubungan kausal secara pasti.

Dalam eksperimen pada penelitian ini, *IndoBERT* memberikan performa keseluruhan terbaik untuk klasifikasi sentimen biner komentar *YouTube* berbahasa Indonesia. Hasil tersebut menunjukkan bahwa kesesuaian bahasa pralatih, formulasi tugas klasifikasi, dan strategi adaptasi model dapat berperan penting pada teks media sosial yang pendek dan informal. Namun, temuan ini tidak membuktikan bahwa *IndoBERT* selalu lebih unggul daripada *DeepSeek* atau *Gemma* pada dataset, *prompt*, metode *fine-tuning*, dan sumber daya komputasi yang berbeda.

Kinerja IndoBERT dalam penelitian ini sejalan secara umum dengan penelitian (Aulia Ruslani & Fauzan, 2025) dan (Manoppo et al., 2025), yang juga

menunjukkan bahwa *IndoBERT* dapat memberikan performa yang kuat pada analisis sentimen berbahasa Indonesia. Namun, *accuracy IndoBERT* sebesar 0,8273 dalam penelitian ini lebih rendah daripada 94,6% yang dilaporkan (Aulia Ruslani & Fauzan, 2025) pada ulasan *e-commerce* dan 84,94% yang dilaporkan (Manoppo et al., 2025) pada data lintas platform mengenai PPN 12%. Perbedaan angka tersebut tidak dapat dibandingkan secara langsung secara kuantitatif, melainkan harus ditinjau lebih dalam dari karakteristik kualitatif datanya. Dataset ulasan *e-commerce* yang digunakan oleh Aulia Ruslani dan Fauzan umumnya memiliki struktur kalimat yang lebih terarah dan spesifik pada fitur produk, sedangkan data lintas platform dari Manoppo mengenai kebijakan publik cenderung menggunakan terminologi yang lebih baku. Sebaliknya, dataset komentar *YouTube* pada topik sosial "Kabur Aja Dulu" dalam penelitian ini sangat sarat akan sarkasme, singkatan yang kompleks, bahasa gaul yang dinamis, serta ekspresi emosi yang ambigu. Karakteristik data yang sangat tidak terstruktur dan informal ini secara inheren menghadirkan tantangan komputasi yang jauh lebih tinggi bagi model untuk memisahkan kelas sentimen secara sempurna. Selain itu, penelitian terdahulu yang ditinjau belum menyediakan pembanding langsung untuk *DeepSeek-R1-Distill-Qwen-1.5B* dan *Gemma-1.1-2B-IT* pada dataset dan prosedur yang sama.

Penelitian ini memiliki beberapa keterbatasan yang perlu dicatat. Pertama, dataset bersifat tidak seimbang dengan dominasi kelas negatif (71,32%), yang dapat memengaruhi kemampuan model dalam mempelajari karakteristik kelas positif secara optimal. Kedua, data hanya bersumber dari dua video *YouTube*, sehingga belum mewakili keseluruhan spektrum opini pada topik “Kabur Aja Dulu” di platform *YouTube*. Ketiga, pelabelan sentimen dilakukan secara otomatis menggunakan pendekatan *lexicon-based labeling* dan hasilnya digunakan sebagai *pseudo-label* referensi tanpa *ground truth* dari annotator manusia independen. Oleh karena itu, nilai akurasi, *precision*, *recall*, dan *F1-score* dalam penelitian ini menunjukkan tingkat kesesuaian prediksi model terhadap *pseudo-label InSet*, bukan ukuran mutlak kesesuaian terhadap penilaian sentimen manusia. Kondisi ini merupakan ancaman validitas karena label berbasis leksikon dapat mengandung *noise* pada komentar yang sarkastik, implisit, ambigu, atau kontekstual. Keempat, *hyperparameter* yang digunakan terbatas pada satu konfigurasi untuk setiap model,

sehingga belum mengeksplorasi seluruh potensi optimal dari masing-masing arsitektur.

Berdasarkan penjabaran metrik evaluasi pada masing-masing kelas, perbedaan karakteristik dan tingkat kemampuan dari ketiga model telah terpetakan dengan jelas. Melalui penilaian performa secara keseluruhan yang komprehensif, *IndoBERT* secara eksplisit ditetapkan sebagai model final untuk diimplementasikan pada tahap penyajian hasil (*deployment*) di dalam aplikasi *Streamlit*. Pemilihan model ini dijustifikasi oleh perolehan kinerja yang paling seimbang, yaitu nilai *Macro-F1* tertinggi sebesar 0,7913 serta kapabilitas terbaik dalam mendeteksi kelas sentimen positif dengan nilai *recall* 0,7186. Seluruh temuan utama dari evaluasi ini, beserta implikasi dari penerapan model terpilih, akan dirangkum ke dalam kesimpulan pada Bab V.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Hasil penelitian ini menunjukkan adanya perbedaan karakteristik dan kemampuan yang signifikan dari ketiga model dalam mengklasifikasikan kelas sentimen, di mana *IndoBERT* terbukti memiliki performa yang paling seimbang. Pada pengenalan kelas positif, *IndoBERT* mengungguli model lainnya dengan perolehan *recall* sebesar 0,7186 dan *F1-score* 0,7047. Di sisi lain, *DeepSeek* menunjukkan kecenderungan yang sangat kuat dan sensitif dalam mendeteksi kelas negatif dengan nilai *recall* tertinggi mencapai 0,9140. Namun, *DeepSeek* memiliki kelemahan dalam mengenali sentimen positif yang ditunjukkan oleh rendahnya nilai *recall* ke angka 0,4070. Sementara itu, *Gemma* menunjukkan performa menengah pada kelas positif dengan *recall* 0,5000, yang lebih baik daripada *DeepSeek* meskipun belum mampu menyamai kinerja *IndoBERT*.

Berdasarkan evaluasi performa secara keseluruhan, *IndoBERT* ditetapkan sebagai model terbaik karena tidak hanya unggul dalam tingkat ketepatan prediksi umum dengan *accuracy* sebesar 0,8273, tetapi juga terbukti paling optimal dalam menangani ketidakseimbangan kelas dengan capaian *macro F1-score* tertinggi yaitu 0,7913. Penentuan metrik evaluasi juga terbukti memengaruhi urutan performa pada dua model generatif. Jika mengacu pada *macro F1-score* sebagai tolak ukur untuk data yang tidak seimbang, *Gemma* menempati peringkat kedua dengan nilai 0,6925, mengungguli *DeepSeek* yang memperoleh 0,6757. Namun, jika evaluasi hanya dititikberatkan pada *accuracy*, posisi *DeepSeek* dengan nilai 0,7687 sedikit menggeser *Gemma* yang memperoleh 0,7620. Secara keseluruhan, pemodelan menggunakan *IndoBERT* memberikan peningkatan metrik performa yang paling komprehensif sekaligus menawarkan keseimbangan klasifikasi terbaik dibandingkan kedua model bahasa besar generasi baru tersebut.

5.2 Saran

Berdasarkan keterbatasan yang ditemukan selama penelitian, beberapa saran untuk penelitian selanjutnya adalah sebagai berikut:

1. Pertama, penggunaan sumber data yang lebih beragam dengan menambahkan lebih banyak video *YouTube* dari berbagai kanal dan perspektif, serta mempertimbangkan teknik balanced sampling yang lebih lanjut atau pengumpulan data yang lebih seimbang sejak awal untuk mengurangi dominasi kelas negatif.

2. Kedua, penelitian selanjutnya sebaiknya membangun data uji independen melalui anotasi manusia oleh sedikitnya dua annotator dengan pedoman pelabelan yang jelas. Kesepakatan antar-annotator dapat diukur, misalnya menggunakan Cohen's kappa, dan perbedaan label diselesaikan melalui adjudikasi. *Pseudo-label* berbasis *InSet* masih dapat dimanfaatkan untuk memperbesar data latih, tetapi evaluasi akhir sebaiknya menggunakan data uji berlabel manusia agar validitas hasil model lebih terjamin

3. Ketiga, eksplorasi *hyperparameter* yang lebih ekstensif perlu dilakukan untuk setiap model, termasuk variasi *learning rate*, *batch size*, jumlah *epoch*, dan konfigurasi *LoRA* (seperti rank dan alpha) untuk menemukan kombinasi yang paling optimal bagi masing-masing arsitektur

4. Keempat, penelitian selanjutnya dapat mempertimbangkan pendekatan *full fine-tuning* pada LLM berukuran kecil seperti *Gemma 2B* sebagai pembanding yang lebih adil, sehingga dapat diketahui apakah keunggulan *IndoBERT* disebabkan oleh arsitektur *encoder-based* atau oleh perbedaan pendekatan *fine-tuning*

5. Kelima, aplikasi web berbasis *Streamlit* yang telah direalisasikan dapat dikembangkan lebih lanjut melalui penambahan kelas netral, visualisasi distribusi sentimen, dukungan berkas yang lebih besar, serta pengujian *usability* dan performa. Pengembangan ini diharapkan memperluas penggunaan aplikasi di luar klasifikasi biner pada teks tunggal dan berkas *CSV*

DAFTAR PUSTAKA

Aditama, F. S., Krismawati, D., & Pramana, S. (2024). MULTICLASS CLASSIFICATION OF MARKETPLACE PRODUCTS WITH MACHINE LEARNING. *Media Statistika*, *17*(1), 25–35. https://doi.org/10.14710/medstat.17.1.25-35

Ahmadian, H., Abidin, T. F., Riza, H., & Muchtar, K. (2023). Transformer-Based Indonesian Language Model for Emotion Classification and Sentiment Analysis. *Proceeding - International Conference on Information Technology and Computing 2023, ICITCOM 2023*. https://doi.org/10.1109/ICITCOM60176.2023.10442970

Alhujaili, R. F., & Yafooz, W. M. S. (2021). Sentiment Analysis for Youtube Videos with User Comments: Review. *Proceedings - International Conference on Artificial Intelligence and Smart Systems, ICAIS 2021*. https://doi.org/10.1109/ICAIS50930.2021.9396049

Angdresey, A., Sitanayah, L., & Tangka, I. L. H. (2025). Sentiment Analysis for Political Debates on YouTube Comments using BERT Labeling, Random Oversampling, and Multinomial Naïve Bayes. *Journal of Computing Theories and Applications*, *2*(3), 342–354. https://doi.org/10.62411/jcta.11668

Appel, G., Grewal, L., Hadi, R., & Stephen, A. T. (2020). The future of social media in marketing. *Journal of the Academy of Marketing Science*, *48*(1). https://doi.org/10.1007/s11747-019-00695-1

Asri, Y., Kuswardani, D., Suliyanti, W. N., Manullang, Y. O., & Ansyari, A. R. (2025). Sentiment analysis based on Indonesian language lexicon and IndoBERT on user reviews PLN mobile application. *Indonesian Journal of Electrical Engineering and Computer Science*, *38*(1), 677. https://doi.org/10.11591/ijeecs.v38.i1.pp677-688

Aulia Ruslani, A. R., & Fauzan, M. D. (2025). Implementasi Model Transformer untuk Analisis Sentimen dan Ringkasan Ulasan E-commerce Berbahasa

Indonesia. *Seminar Nasional Inovasi Vokasi*. https://prosiding.pnj.ac.id/index.php/sniv/article/view/4188

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*. https://doi.org/10.1016/j.knosys.2021.107134

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., … Zhang, Z. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *ArXiv*, arXiv:2501.12948. https://arxiv.org/abs/2501.12948

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. *ArXiv*. http://arxiv.org/abs/2305.14314

El Azzouzy, O., Chanyour, T., & Andaloussi, S. J. (2025). Transformer-based models for sentiment analysis of YouTube video comments. *Scientific African*, *29*. https://doi.org/10.1016/j.sciaf.2025.e02836

Fatani, I. I., & Irawan, H. (2024). Twitter, Instagram, Youtube Speak: Understanding Sentiments on LRT Jabodebek Services via Inset Lexicon, IndoBERT and BERTopic Approaches. *J. Electrical Systems*, *20*(4), 1028–1035.

Gadyanavar, P., Laddi, M., Jadhav, P., Katti, V. S., Pujari, J., & Jamadar, A. J. (2025). *Sentiment Analysis of YouTube Comments Using Bidirectional Encoder Representations from Transformers Neural Network Model*. 683–689. https://doi.org/10.5220/0013640000004664

Gemma Team. (2024). Gemma: Open Models Based on Gemini Research and Technology. *ArXiv*, arXiv:2403.08295. https://arxiv.org/abs/2403.08295

Gosztonyi, G., Bálint, J., & Lendvai, G. F. (2025). Public Figures and Social Media from a Freedom of Expression Viewpoint in the Recent U.S. and EU

Jurisdiction. *Journalism and Media*, *6*(1). https://doi.org/10.3390/journalmedia6010026

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: an Overview*. http://arxiv.org/abs/2008.05756

Jaya Kusoema, W., & Ibrahim, I. (2025). Analisis Sentimen dalam Kasus Korupsi PT. Pertamina menggunakan Metode indoBERT dan RCNN. *Sistemasi: Jurnal Sistem Informasi*, *14*. http://sistemasi.ftik.unisi.ac.id

Koto, F., & Rahmaningtyas, G. Y. (2017). Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs. *Proceedings of the 2017 International Conference on Asian Language Processing, IALP 2017*, *2018-January*. https://doi.org/10.1109/IALP.2017.8300625

Krishnappa K, C. T., Nayaka, N. B., C, P. B., & V, S. U. (2025). Multilingual Sentimental Analysis of Youtube Comments. *International Journal for Multidisciplinary Research (IJFMR)*, *7*(1). www.ijfmr.com

Lin, C. H., & Nuha, U. (2023). Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy. *Journal of Big Data*, *10*(1). https://doi.org/10.1186/s40537-023-00782-9

Ma'aly, A. N., Pramesti, D., Fathurahman, A. D., & Fakhrurroja, H. (2024). Exploring Sentiment Analysis for the Indonesian Presidential Election Through Online Reviews Using Multi-Label Classification with a Deep Learning Algorithm. *Information*, *15*(11). https://doi.org/10.3390/info15110705

Manoppo, M. R., Kolang, I. C., Nur Fiat, D. N., Mawara, R. M. C., Sumarno, A. D. P., Yusupa, A., & Tarigan, V. (2025). ANALISIS SENTIMEN PUBLIK DI MEDIA SOSIAL TERHADAP KENAIKAN PPN 12% DI INDONESIA MENGGUNAKAN INDOBERT. *Jurnal Kecerdasan Buatan Dan Teknologi Informasi*, *4*(2), 152–163. https://doi.org/10.69916/jkbti.v4i2.322

Otter, D. W., Medina, J. R., & Kalita, J. K. (2021). A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural*

Networks and Learning Systems, *32*(2). https://doi.org/10.1109/TNNLS.2020.2979670

Pandey, S., & Jain, M. (2021). Time Series Visualization using Transformer for Prediction of Natural Catastrophe. *International Journal of Science and Research (IJSR)*, *10*(10), 1137–1146. https://doi.org/10.21275/sr211022155010

Priantoro, S. D. B., Rohman, M. G., & Zamroni, M. R. (2025). KLASIFIKASI KUALITAS UDARA DENGAN METODE NAIVE BAYES BERBASIS WEB. *Rabit : Jurnal Teknologi Dan Sistem Informasi Univrab*, *10*(2), 1024–1035. https://doi.org/10.36341/rabit.v10i2.6447

Puspitasari, A., Paradhita, A. N., Tineka, Y. W., Sulistyowati, V., Noriska, N. K. S., & Haryanto. (2024). Natural Language Processing (NLP) Technology for Chatbot Website. *Jurnal Penelitian Pendidikan IPA*, *10*(SpecialIssue), 319–324. https://doi.org/10.29303/jppipa.v10ispecialissue.8241

Ranjan, N., Mundada, K., Phaltane, K., & Ahmad, S. (2016). A Survey on Techniques in NLP. *International Journal of Computer Applications*, *134*(8). https://doi.org/10.5120/ijca2016907355

Riyadi, S., Salsabila, L. K., Damarjati, C., & Karim, R. A. (2024). Sentiment Analysis of YouTube Users on Blackpink Kpop Group Using IndoBERT. *INTENSIF: Jurnal Ilmiah Penelitian Dan Penerapan Teknologi Sistem Informasi*, *8*(2), 2549–6824. https://doi.org/10.29407/intensif.v8n2.22678

Rizkia, A. S., Wufron, W., & Roji, F. F. (2025). Analisis Sentimen Coretax: Perbandingan Pelabelan Data Manual, Transformers-Based, dan Lexicon-Based pada Performa IndoBERT. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *5*(3). https://doi.org/10.57152/malcom.v5i3.2151

Salma, T. D., Kurniawan, M. F., Darmawan, R., & Basri, A. (2025). Analisis Sentimen Berbasis Transformer: Persepsi Publik terhadap Nusantara pada Perayaan Kemerdekaan Indonesia yang Pertama. *Jurnal JTIK (Jurnal*

Teknologi Informasi Dan Komunikasi), *9*(2). https://doi.org/10.35870/jtik.v9i2.3535

Sari, P., Silaban, M. J., Mirza, D., Nafilah, N., Tanjung, S. Z., Pancasila, P., Bahasa, F., & Seni, D. (2025). Menghadapi Ancaman Nasionalisme Disintegrasi Bangsa di Tengah Trend Kabur Aja Dulu. *Jurnal Bintang Pendidikan Indonesia*, *3*(2), 193–199. https://doi.org/10.55606/jubpi.v3i2.3821

Shivsharan, N., Kambli, V., Dabholkar, S., Dalvi, A., & Sukali, T. (2025). Evaluating Machine Learning Models for Sentiment Analysis of YouTube Comments and Creating an Accessible Web Application for Comment Analysis. *Procedia Computer Science*, *258*, 3165–3174. https://doi.org/10.1016/j.procs.2025.04.574

Suhendar, H., Slamet, C., & Syaripudin, U. (2025). Analisis Sentimen Hasil Transkripsi Audio Berbahasa Indonesia Menggunakan T5 (Text-to-Text Transfer Transformer). *SMATIKA JURNAL*, *15*(01). https://doi.org/10.32664/smatika.v15i01.1521

Supriyono, Wibawa, A. P., Suyono, & Kurniawan, F. (2024). Advancements in natural language processing: Implications, challenges, and future directions. *Telematics and Informatics Reports*, *16*. https://doi.org/10.1016/j.teler.2024.100173

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-December*.

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL-IJCNLP 2020*. https://doi.org/10.18653/v1/2020.aacl-main.85

Yadalam, P. K., Thaha, M., Natarajan, P. M., & Ardila, C. M. (2025). An explainable RoBERTa approach to analyzing panic and anxiety sentiment in oral health education YouTube comments. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-06560-2

Yang, A., Zhang, B., Hui, B., Gao, B., Yu, B., Li, C., Liu, D., Tu, J., Zhou, J., Lin, J., Lu, K., Xue, M., Lin, R., Liu, T., Ren, X., & Zhang, Z. (2024). Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement. *ArXiv*. http://arxiv.org/abs/2409.12122

Yudertha, A., & Dwimalida Putri, R. (2025). Pemetaan Tren Machine Learning dalam Penelitian Ilmu Kimia menggunakan LLM dengan Multi-Turn Prompting Mapping Machine Learning Trends in Chemistry Research using LLM with Multi-Turn Prompting. *Sistemasi: Jurnal Sistem Informasi*, *14*(2). http://sistemasi.ftik.unisi.ac.id