KLASIFIKASI SENTIMEN KOMENTAR PUBLIK DI YOUTUBE TENTANG KECERDASAN BUATAN (AI) DALAM PENDIDIKAN MENGGUNAKAN BERT DAN INDOROBERTA

TESIS

Oleh :

MUHAMMAD MUTAWAKKIL ALLALLAH

NIM. 240605220005

MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2026

KLASIFIKASI SENTIMEN KOMENTAR PUBLIK DI YOUTUBE TENTANG KECERDASAN BUATAN (AI) DALAM PENDIDIKAN MENGGUNAKAN BERT DAN INDOROBERTA

TESIS

Diajukan Kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Magister Komputer (M.Kom)

Oleh:
MUHAMMAD MUTAWAKKIL ALALLAH
NIM. 240605220005

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

KLASIFIKASI SENTIMEN KOMENTAR PUBLIK DI YOUTUBE TENTANG KECERDASAN BUATAN (AI) DALAM PENDIDIKAN MENGGUNAKAN BERT DAN INDOROBERTA

TESIS

Oleh:
MUHAMMAD MUTAWAKKIL ALALLAH
NIM. 240605220005

Telah diperiksa dan disetujui untuk diuji:
Tanggal: 28 April 2026

Pembimbing I,	Pembimbing II,
Dr. Ir. Mokhamad Amin Hariyadi, M.T	Prof. Dr. Triyo Supriyatno, M.Ag.
NIP. 19670118 200501 1 001	NIP. 19700427 200003 1 001

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

Prof. Dr. Ir. Muhammad Faisal, M.T
NIP. 19740510 200501 1 007

KLASIFIKASI SENTIMEN KOMENTAR PUBLIK DI YOUTUBE TENTANG KECERDASAN BUATAN (AI) DALAM PENDIDIKAN MENGGUNAKAN BERT DAN INDOROBERTA

TESIS

Oleh:
MUHAMMAD MUTAWAKKIL ALALLAH
NIM. 240605220005

Telah Dipertahankan di Depan Dewan Penguji Tesis
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Magister Komputer (M.Kom)
Tanggal: 28 Mei 2026

Susunan Dewan Penguji			**Tanda Tangan**
Penguji I	:	Prof. Dr. Hj. Sri Harini, M.Si NIP. 19731014 200112 2 002	( )
Penguji II	:	Dr. Irwan Budi Santoso, M.Kom NIP. 19770103 201101 1 004	( )
Pembimbing I	:	Dr. Ir. Mokhamad Amin Hariyadi, M.T NIP. 19670118 200501 1 001	( )
Pembimbing II	:	Prof. Dr. Triyo Supriyatno, M.Ag. NIP. 19700427 200003 1 001	( )

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

Prof. Dr. Ir. Muhammad Faisal, M.T
NIP. 19740510 200501 1 007

HALAMAN PERNYATAAN

Saya yang bertanda tangan di bawah ini:

Nama	: Muhammad Mutawakkil Alallah
NIM	: 240605220005
Program Studi	: Magister Informatika
Fakultas	: Sains dan Teknologi

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

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

Malang, 21 April 2026
Yang membuat pernyataan

Muhammad Mutawakkil Alallah
NIM. 240605220005

MOTTO

(مَنْ جَدَّ وَجَدَ)

"Barang siapa bersungguh-sungguh, maka ia akan memperoleh apa yang diusahakannya."

(Al-Hasyimi, 1996).

PERSEMBAHAN

Dengan mengucapkan syukur kehadirat Allah SWT, *Alhamdulillāhi Rabbil ʻĀlamīn*, tesis ini penulis persembahkan kepada:

1. Kedua orang tua tercinta, Aba Abdus Somad dan Umi Ismawati, yang senantiasa memberikan doa, kasih sayang, dukungan, serta semangat yang tiada henti kepada penulis.
2. Seluruh keluarga besar penulis, yaitu Alm. Mba Syafi'i, Alm. Mbah Nidin, Almh. Mbah Ya, Mbah Bunari, Adek Annisaul Marhamah, Adek Indra Rosyidah, serta keluarga mertua: Bapak Moh. Rasyid, Umi Khusnul Khotimah, Bapak, Mama, Almh. Desi Humairoh, Kong, Uti, Adek Rani, Adek Rara, De San, De Ririn, Dek Ifroh, Dek Serli, De Sutres Lakek, dan Binik De Cip, yang senantiasa memberikan doa, perhatian, serta dukungan moral maupun spiritual.
3. Seluruh Civitas Akademika Universitas Islam Negeri Maulana Malik Ibrahim Malang yang telah memberikan kesempatan kepada penulis untuk menimba ilmu pengetahuan, teknologi, dan nilai-nilai keislaman.
4. Seluruh rekan-rekan Asosiasi Mahasiswa Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang dari seluruh angkatan atas kebersamaan, kerja sama, dan dukungan selama masa perkuliahan.
5. Teman-teman santri PPIQ Tahfidzul Qur'an PPIQ Tahfidzul Qur'an Nurul Jadid yang senantiasa memberikan motivasi dan semangat kepada penulis dalam menyelesaikan tesis ini.
6. Seluruh rekan-rekan GEMMATAN (Generasi Muda Mahasiswa Ahlith Thoriqoh Al-Mu'tabaroh An-Nahdliyyah) yang telah memberikan dukungan, kebersamaan, motivasi, serta pengalaman organisasi yang bermakna selama proses penyelesaian studi dan tesis ini.
7. Seluruh pihak, baik keluarga, sahabat, maupun rekan-rekan lainnya yang tidak dapat disebutkan satu per satu, yang telah memberikan dukungan, bantuan, doa, dan semangat dalam proses penyelesaian tesis ini.

KATA PENGANTAR

Assalamu'alaikum Wr. Wb.

Puji syukur kehadirat Allah SWT yang telah melimpahkan rahmat, taufik, dan hidayah-Nya, sehingga penulis dapat menyelesaikan studi pada Program Studi Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang serta menyelesaikan tesis ini dengan baik.

Shalawat serta salam semoga senantiasa tercurah kepada junjungan Nabi Muhammad SAW yang telah membimbing umat manusia menuju jalan kebenaran dan ilmu pengetahuan.

Selanjutnya, penulis menyampaikan ucapan terima kasih disertai doa *jazākumullāhu aḥsanal jazā'* kepada seluruh pihak yang telah membantu dalam penyelesaian tesis ini. Ucapan terima kasih penulis sampaikan kepada:

1. Rektor Prof. Dr. Hj. Ilfi Nur Diana, M.Si. yang telah memberikan dukungan dan fasilitas selama penulis menempuh studi.
2. Dekan Fakultas Sains dan Teknologi Dr. Agus Mulyono, M.Kes. yang telah memberikan arahan dan dukungan akademik kepada penulis.
3. Ketua Program Studi Magister Informatika Prof. Dr. Moh. Faisal, M.T. yang telah memberikan motivasi dan pelayanan akademik selama proses studi.
4. Bapak Dr. Ir. Mokhamad Amin Hariyadi, M.T., M.Kom. dan Bapak Prof. Dr. Triyo Supriyatno, M.Ag. selaku dosen pembimbing tesis yang telah banyak memberikan arahan, masukan, serta pengalaman akademik yang sangat berharga kepada penulis.
5. Segenap sivitas akademika Program Studi Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang, khususnya seluruh Bapak/Ibu dosen, atas ilmu, bimbingan, dan motivasi yang telah diberikan selama masa studi.

6. Aba tercinta Abdus Shomad dan Umi tercinta Ismawati yang senantiasa memberikan doa, kasih sayang, dukungan, serta restu kepada penulis dalam menuntut ilmu dan menyelesaikan tesis ini.
7. Pengasuh dan Dewan Pengasuh Pondok Pesantren Nurul Jadid yang telah memberikan ilmu, bimbingan, dan keteladanan kepada penulis.
8. Semua pihak yang telah membantu dalam penyusunan dan penyelesaian tesis ini yang tidak dapat penulis sebutkan satu per satu.

Penulis menyadari bahwa tesis ini masih memiliki keterbatasan dan kekurangan. Oleh karena itu, penulis mengharapkan kritik dan saran yang konstruktif demi penyempurnaan karya ilmiah ini. Penulis berharap semoga tesis ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan, khususnya di bidang informatika dan kecerdasan buatan.

Wassalamu'alaikum Wr. Wb.

Malang, 21 April 2026

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Alallah, Muhammad Mutawakkil. 2026. **KLASIFIKASI SENTIMEN KOMENTAR PUBLIK DI YOUTUBE TENTANG KECERDASAN BUATAN (AI) DALAM PENDIDIKAN MENGGUNAKAN BERT DAN INDOROBERTA.** Program Studi Magister Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Dr. Ir. Mokhamad Amin Hariyadi, M.T (II) Prof. Dr. Triyo Supriyatno, M.Ag.

Kata Kunci: Klasifikasi Sentimen, BERT, IndoRoBERTa, YouTube, Kecerdasan Buatan

Meningkatnya penggunaan media sosial dalam penyampaian opini publik terkait kecerdasan buatan (AI) di bidang pendidikan mendorong kebutuhan akan metode otomatis untuk klasifikasi sentimen yang lebih sistematis dan akurat. Penelitian ini bertujuan untuk mengklasifikasikan sentimen komentar publik berbahasa Indonesia pada platform YouTube ke dalam tiga kategori, yaitu positif, netral, dan negatif, menggunakan model BERT dan IndoRoBERTa. Data dikumpulkan menggunakan *YouTube Data API* dengan pendekatan *keyword-based crawling* pada periode Agustus 2020 hingga Desember 2025. Setelah proses validasi, dilakukan pelabelan manual terhadap 10.834 komentar yang berasal dari 167 video dan 145 kanal YouTube. Tahap *pre-processing* meliputi *case folding*, *cleaning*, *stopword removal*, penghapusan data duplikat, *filtering* topik, serta penanganan data teks sehingga diperoleh 4.726 data yang digunakan dalam eksperimen. Proses *tokenization* dilakukan menggunakan *WordPiece Tokenizer* pada BERT dan *Byte Pair Encoding (BPE)* pada IndoRoBERTa untuk menyesuaikan representasi teks masing-masing model. Pengujian dilakukan melalui skenario *baseline* tanpa *fine-tuning* dan skenario *fine-tuning* dengan pembagian data 70:15:15, 80:10:10, dan 90:5:5 serta penerapan *class weight* untuk mengatasi ketidakseimbangan kelas. Hasil penelitian menunjukkan bahwa kedua model mengalami peningkatan performa setelah *fine-tuning*, dengan IndoRoBERTa secara konsisten memberikan hasil yang lebih baik dibandingkan BERT pada sebagian besar skenario pengujian. Secara keseluruhan, penelitian ini menunjukkan bahwa pendekatan berbasis *Transformer* efektif digunakan untuk klasifikasi sentimen komentar publik berbahasa Indonesia pada topik kecerdasan buatan dalam pendidikan.

ABSTRACT

Alallah, Muhammad Mutawakkil. 2026. Sentiment Classification of Public Comments on YouTube Regarding Artificial Intelligence (AI) in Education Using BERT and IndoRoBERTa. Master's Program in Informatics, Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang. Supervisors: : (I) Dr. Ir. Mokhamad Amin Hariyadi, M.T (II) Prof. Dr. Triyo Supriyatno, M.Ag.

Keywords: Sentiment Classification, BERT, IndoRoBERTa, YouTube, Artificial Intelligence

The increasing use of social media as a medium for expressing public opinions regarding Artificial Intelligence (AI) in education has created a need for systematic and accurate automated sentiment classification methods. This study aims to classify public comments in Indonesian language on YouTube into three categories, namely positive, neutral, and negative, using BERT and IndoRoBERTa models. Data were collected using the YouTube Data API through a keyword-based crawling approach from August 2020 to December 2025. After validation, manual labeling was conducted on 10,834 comments obtained from 167 videos and 145 YouTube channels. The preprocessing stage included case folding, cleaning, stopword removal, duplicate elimination, topic filtering, and text handling, resulting in 4,726 data samples used for experiments. Tokenization was performed using WordPiece Tokenizer for BERT and Byte Pair Encoding (BPE) for IndoRoBERTa to adapt text representation to each model. Experiments were conducted using a baseline scenario without fine-tuning and fine-tuning scenarios with data splits of 70:15:15, 80:10:10, and 90:5:5, along with class weight implementation to address class imbalance. The results show that both models improved after fine-tuning, with IndoRoBERTa consistently outperforming BERT in most experimental settings. Overall, the study demonstrates that Transformer-based approaches are effective for sentiment classification of Indonesian-language public comments on AI in education.

الملخص

ألاله، محمد متوكل. 2026. تصنيف المشاعر لتعليقات الجمهور على يوتيوب حول الذكاء الاصطناعي **(AI)** في التعليم باستخدام نموذجي **BERT** و **.IndoRoBERTa** برنامج ماجستير المعلوماتية، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم الإسلامية الحكومية مالانج. المشرفان: (1) د. المهندس محمد أمين هريادي، (2) الأستاذ الدكتور تريو سوبرياتنو، ماجستير في الدراسات الإسلامية.

الكلمات المفتاحية: تصنيف المشاعر، BERT، IndoRoBERTa، **يوتيوب**، الذكاء الاصطناعي

إن الاستخدام المتزايد لوسائل التواصل الاجتماعي كوسيلة للتعبير عن آراء الجمهور حول الذكاء الاصطناعي في مجال التعليم أدى إلى الحاجة إلى أساليب آلية دقيقة ومنهجية لتصنيف المشاعر. تهدف هذه الدراسة إلى تصنيف تعليقات الجمهور باللغة الإندونيسية على منصة يوتيوب إلى ثلاث فئات هي: إيجابية ومحايدة وسلبية باستخدام نموذجي BERT و .IndoRoBERTa تم جمع البيانات باستخدام واجهة برمجة تطبيقات يوتيوب (YouTube Data API) من خلال أسلوب الزحف المعتمد على الكلمات المفتاحية خلال الفترة من أغسطس 2020 إلى ديسمبر 2025. وبعد عملية التحقق، تم إجراء التصنيف اليدوي على 10,834 تعليقاً مأخوذة من 167 فيديو و145 قناة على يوتيوب. شملت مرحلة المعالجة المسبقة تطبيع النصوص(case folding) ، والتنظيف، وإزالة الكلمات الشائعة، وحذف البيانات المكررة، وتصفية الموضوع، ومعالجة النصوص، مما أدى إلى الحصول على 4,726 عينة مستخدمة في التجارب. تم تنفيذ عملية الترميز باستخدام WordPiece Tokenizer في نموذج BERT و Byte Pair Encoding (BPE)في نموذج IndoRoBERTa لملاءمة تمثيل النص مع كل نموذج. أُجريت التجارب باستخدام سيناريو أساسي بدون ضبط دقيق (fine-tuning) وسيناريوهات مع الضبط الدقيق بتقسيمات بيانات 70:15:15 و80:10:10 و90:5:5، مع استخدام أوزان الفئات لمعالجة عدم توازن البيانات. أظهرت النتائج أن كلا النموذجين حققا تحسناً بعد الضبط الدقيق، مع تفوق IndoRoBERTa بشكل مستمر على BERTفي معظم الإعدادات التجريبية. وبشكل عام، تثبت الدراسة أن الأساليب المعتمدة على نماذج Transformerفعالة في تصنيف مشاعر تعليقات الجمهور باللغة الإندونيسية حول الذكاء الاصطناعي في التعليم.

BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi kecerdasan buatan (AI) dalam beberapa tahun terakhir, khususnya pada periode 2021–2025, menunjukkan peningkatan yang pesat dengan penerapan di berbagai sektor seperti industri, kesehatan, pendidikan, dan hiburan (Chew & Achananuparp, 2022; Sahar & Munawaroh, 2025). AI berperan penting dalam otomatisasi, analisis data berskala besar, serta pengambilan keputusan prediktif, yang menjadi pilar utama Revolusi Industri 4.0 (Jamil *et al*., 2024). Teknologi ini menawarkan efisiensi dan produktivitas tinggi, namun juga menimbulkan tantangan baru terkait etika, sosial, dan psikologis, termasuk kekhawatiran masyarakat mengenai hilangnya lapangan kerja, penyalahgunaan data, dan dominasi mesin dalam kehidupan sehari-hari *(*Bello *et al*., 2023).

Seiring meningkatnya perhatian publik terhadap AI, platform digital seperti YouTube berperan penting dalam membentuk opini dan persepsi masyarakat. Sebagai salah satu platform berbasis video terbesar di dunia, YouTube memiliki jutaan saluran aktif dan memuat ratusan ribu unggahan setiap harinya (Raza *et al*., 2024). Interaksi pengguna melalui kolom komentar menjadi sumber data yang kaya untuk memahami pandangan dan emosi publik, termasuk terkait penerapan AI di bidang pendidikan. Melalui komentar-komentar tersebut, masyarakat mengekspresikan tanggapan beragam, mulai dari apresiasi atas kemajuan teknologi hingga kekhawatiran terhadap dampak sosialnya. Klasifikasi sentimen terhadap data ini menjadi relevan untuk memahami opini publik secara empiris dan memberikan arahan bagi pengembangan AI yang selaras dengan kebutuhan sosial.

Beberapa penelitian terdahulu telah meneliti fenomena serupa. Jamil *et al*. (2024) mengklasifikasikan komentar publik pada video simulasi bencana di YouTube menggunakan Naïve Bayes dan SVM dengan representasi TF-IDF, sementara Nahid *et al*. (2024) menganalisis emosi publik pada isu politik

Indonesia 2024 menggunakan CNN dan Bi-LSTM. Xu *et al*. (2025) meneliti persepsi publik terhadap teknologi deepfake di Reddit, Alawi & Bozkurt (2024) mengkaji opini publik terhadap universitas melalui Twitter, dan Bello *et al*. (2023) mengembangkan sistem klasifikasi tweet menggunakan model BERT. Namun, sebagian besar penelitian terdahulu tidak menelaah persepsi publik terhadap AI di bidang pendidikan, terutama komentar YouTube berbahasa Indonesia, serta jarang mengaitkan hasil analisis dengan nilai sosial dan etika Islam.

Dalam perspektif Islam, kemajuan teknologi termasuk kecerdasan buatan (AI) dipandang sebagai manifestasi dari anugerah akal dan ilmu pengetahuan yang diberikan Allah SWT kepada manusia. Hal ini ditegaskan dalam QS. Al-Baqarah [2]: 31:

وَعَلَّمَ اٰدَمَ الْاَسْمَاۤءَ كُلَّهَا ثُمَّ عَرَضَهُمْ عَلَى الْمَلٰۤىِٕكَةِ فَقَالَ اَنْۢبِـُٔوْنِيْ بِاَسْمَاۤءِ هٰٓؤُلَاۤءِ اِنْ كُنْتُمْ صٰدِقِيْنَ ﴿٣١﴾

Artinya: *"Dia mengajarkan kepada Adam nama-nama (benda) seluruhnya, kemudian Dia memperlihatkannya kepada para malaikat, seraya berfirman, 'Sebutkan kepada-Ku nama-nama (benda) ini jika kamu benar!'"* (QS. Al-Baqarah [2]: 31); terjemahan Qur'an NU Online, 2026).

Menurut Tafsir al-Munir, ayat ini menunjukkan keistimewaan manusia dalam penguasaan ilmu sebagai dasar pengembangan pengetahuan dan teknologi, termasuk AI yang merepresentasikan kemampuan berpikir dan pengambilan keputusan rasional (Alfiansyah et al., 2023). Hal ini diperkuat oleh QS. Al-'Alaq [1–5]:

اقْرَأْ بِاسْمِ رَبِّكَ الَّذِي خَلَقَ (١) خَلَقَ الْإِنسَانَ مِنْ عَلَقٍ (٢) اقْرَأْ وَرَبُّكَ الْأَكْرَمُ (٣) الَّذِي عَلَّمَ بِالْقَلَمِ (٤) عَلَّمَ الْإِنسَانَ مَا لَمْ يَعْلَمْ (٥)

Artinya: *"Bacalah dengan (menyebut) nama Tuhanmu yang menciptakan!. Dia menciptakan manusia dari segumpal darah. Bacalah! Tuhanmulah Yang Mahamulia. Yang mengajar (manusia) dengan pena. Dia mengajarkan manusia apa yang tidak diketahuinya."* (QS. Al-'Alaq [96]: 1–5; terjemahan Qur'an NU Online, 2026).

Ayat ini menegaskan bahwa pengembangan teknologi harus berbasis ilmu, etika, dan kemaslahatan, sejalan dengan *maqāṣid al-sharīʿah* yang menempatkan perlindungan akal dan kesejahteraan manusia sebagai dasar moral penggunaan AI

(Masykur & Solekhah, 2021). Prinsip kehati-hatian dalam penggunaan informasi juga ditegaskan dalam QS. Al-Isra' [17]: 36:

وَلَا تَقْفُ مَا لَيْسَ لَكَ بِهٖ عِلْمٌۗ اِنَّ السَّمْعَ وَالْبَصَرَ وَالْفُؤَادَ كُلُّ اُولٰۤىِٕكَ كَانَ عَنْهُ مَسْـُٔوْلًا ﴿٣٦﴾

Artinya: *"Janganlah engkau mengikuti sesuatu yang tidak kauketahui. Sesungguhnya pendengaran, penglihatan, dan hati nurani, semua itu akan diminta pertanggungjawabannya."* (QS. Al-Isra' [17]: 36; terjemahan Qur'an NU Online, 2026).

Ayat ini menegaskan pentingnya verifikasi dan tanggung jawab dalam penggunaan informasi, yang relevan dengan etika penggunaan AI (Ali et al., 2025). Hal ini diperkuat pula dalam QS. Al-Hujurat [49]: 6:

يٰٓاَيُّهَا الَّذِيْنَ اٰمَنُوْٓا اِنْ جَاۤءَكُمْ فَاسِقٌۢ بِنَبَاٍ فَتَبَيَّنُوْٓا اَنْ تُصِيْبُوْا قَوْمًاۢ بِجَهَالَةٍ فَتُصْبِحُوْا عَلٰى مَا فَعَلْتُمْ نٰدِمِيْنَ ﴿٦﴾

Artinya: *"Wahai orang-orang yang beriman, jika seorang fasik datang kepadamu membawa berita penting, maka telitilah kebenarannya agar kamu tidak mencelakakan suatu kaum karena ketidaktahuan(-mu) yang berakibat kamu menyesali perbuatanmu itu."* (QS. Al-Hujurat [49]: 6; terjemahan Qur'an NU Online, 2026).

Prinsip ini relevan dengan isu verifikasi data, misinformasi, dan etika AI dalam konteks digital (Molla & Ahsan, 2025). Selain itu, dalam hadis riwayat Muslim dinyatakan:

عَنْ أَبِي هُرَيْرَةَ أَنَّ رَسُولَ اللهِ ﷺ قَالَ مَنْ دَعَا إِلَى هُدًى كَانَ لَهُ مِنَ الأَجْرِ مِثْلُ أُجُورِ مَنْ تَبِعَهُ لاَ يَنْقُصُ ذَلِكَ مِنْ أُجُورِهِمْ شَيْئًا وَمَنْ دَعَا إِلَى ضَلاَلَةٍ كَانَ عَلَيْهِ مِنَ الإِثْمِ مِثْلُ آثَامِ مَنْ تَبِعَهُ لاَ يَنْقُصُ ذَلِكَ مِنْ آثَامِهِمْ شَيْئًا

Artinya: *"Barang siapa yang menunjukkan kepada kebaikan, maka ia mendapatkan pahala seperti orang yang melakukannya."* (HR. Muslim).

Hadis ini menegaskan bahwa pemanfaatan teknologi bernilai ibadah apabila diarahkan pada kemaslahatan. Dalam kerangka etika Islam, penggunaan AI harus berorientasi pada kemaslahatan dan pencegahan kerusakan, sesuai kaidah fiqh *"Dar'u al-mafāsid muqaddam 'alā jalb al-maṣāliḥ"* serta prinsip maqāṣid al-sharī'ah (Ali *et al*., 2025).

Dari sisi akademis, penelitian ini berkontribusi pada pengembangan ilmu Informatika, khususnya *Natural Language Processing*, *Deep Learning*, dan analisis data sosial. Secara sosial, penelitian ini membantu memahami respons masyarakat terhadap inovasi AI di ruang digital. Dari perspektif keislaman,

penelitian ini merefleksikan tanggung jawab manusia sebagai *khalifah* dalam mengelola ilmu pengetahuan dan teknologi secara bijak, adil, dan bertanggung jawab, sebagaimana ditegaskan dalam QS. Al-Baqarah [2]: 30:

وَاِذْ قَالَ رَبُّكَ لِلْمَلٰۤىِٕكَةِ اِنِّيْ جَاعِلٌ فِى الْاَرْضِ خَلِيْفَةً ۗقَالُوْٓا اَتَجْعَلُ فِيْهَا مَنْ يُّفْسِدُ فِيْهَا وَيَسْفِكُ الدِّمَاۤءَ ۚوَنَحْنُ نُسَبِّحُ بِحَمْدِكَ وَنُقَدِّسُ لَكَ ۗقَالَ اِنِّيْٓ اَعْلَمُ مَا لَا تَعْلَمُوْنَ ﴿٣٠﴾

Artinya: *"(Ingatlah) ketika Tuhanmu berfirman kepada para malaikat, "Aku hendak menjadikan khalifah di bumi." Mereka berkata, "Apakah Engkau hendak menjadikan orang yang merusak dan menumpahkan darah di sana, sedangkan kami bertasbih memuji-Mu dan menyucikan nama-Mu?" Dia berfirman, "Sesungguhnya Aku mengetahui apa yang tidak kamu ketahui."* (QS. Al-Baqarah [2]: 30; terjemahan Qur'an NU Online, 2026).

Ayat ini menegaskan mandat manusia sebagai *khalifah* yang bertanggung jawab dalam mengelola bumi, termasuk dalam pengembangan dan pemanfaatan AI agar selaras dengan nilai kemaslahatan, keadilan, dan etika kemanusiaan. Prinsip ini diperkuat oleh kajian kontemporer yang menempatkan *maqāṣid al-sharīʿah* seperti perlindungan akal *(ḥifẓ al-ʿaql)*, jiwa *(ḥifẓ al-nafs)*, dan kemaslahatan umum sebagai dasar etika dalam tata kelola AI yang bertanggung jawab dan bermartabat (Muchtasor, 2025).

Kebaruan penelitian terletak pada fokus analisis persepsi publik terhadap AI di bidang pendidikan melalui komentar YouTube berbahasa Indonesia, serta pengintegrasian nilai etika dan prinsip Islam, seperti keadilan, verifikasi informasi, dan *maqāṣid al-sharīʿah*. Pendekatan teknis menggunakan model IndoRoBERTa memungkinkan interpretasi kontekstual komentar secara lebih akurat dan efisien dibanding metode sebelumnya, sehingga penelitian ini menghadirkan kontribusi ilmiah, sosial, dan religius yang komprehensif.

Seiring dengan berkembangnya pendekatan analisis sentimen berbasis pembelajaran mendalam, model *Transformer* seperti BERT dan variannya terbukti mampu menangkap konteks semantik secara lebih akurat dibandingkan metode klasik berbasis fitur statistik. BERT *(Bidirectional Encoder Representations from Transformers)* memiliki keunggulan dalam memahami hubungan dua arah antar kata dalam suatu kalimat, sehingga efektif untuk menganalisis teks yang bersifat kontekstual dan ambigu, seperti komentar media sosial yang mengandung variasi

makna dan ekspresi subjektif (Bello *et al*., 2023). Namun, karena BERT dilatih menggunakan korpus berbahasa Inggris, penerapannya pada teks berbahasa Indonesia terutama pada komentar YouTube yang cenderung informal, tidak baku, dan kaya variasi linguistik memerlukan adaptasi bahasa. Oleh karena itu, penelitian ini menggunakan IndoRoBERTa, yaitu pengembangan RoBERTa yang telah dilatih khusus pada korpus besar bahasa Indonesia, sehingga lebih representatif dalam menangkap karakteristik bahasa lokal, ragam informal, serta pola ekspresi pengguna YouTube Indonesia. Perbandingan penggunaan BERT dan IndoRoBERTa dalam penelitian ini bertujuan untuk mengevaluasi efektivitas model pralatih umum dan model spesifik bahasa dalam mengklasifikasikan sentimen publik terhadap AI di bidang pendidikan secara lebih akurat dan kontekstual.

1.2 Pernyataan Masalah

Pernyataan masalah dalam penelitian ini adalah seberapa besar kontribusi performa model BERT dan IndoRoBERTa dalam mengklasifikasikan komentar publik berbahasa Indonesia di platform YouTube mengenai kecerdasan buatan (AI) dalam bidang pendidikan?

1.3 Tujuan Penelitian

Penelitian ini bertujuan untuk mengukur kontribusi performa model BERT dan IndoRoBERTa dalam mengklasifikasikan komentar publik berbahasa Indonesia di platform YouTube mengenai kecerdasan buatan (AI) dalam bidang pendidikan.

1.4 Batasan Masalah

Penelitian ini dibatasi agar fokus pada hal-hal berikut:

1. Data yang dianalisis berupa komentar YouTube berbahasa Indonesia yang berkaitan dengan topik kecerdasan buatan (AI) dalam konteks pendidikan
2. Klasifikasi sentimen dalam penelitian ini dibatasi pada tiga kategori, yaitu positif, negatif, dan netral. Pembatasan ini diterapkan untuk menjaga

konsistensi dengan pendekatan analisis sentimen dasar (*polarity-based sentiment analysis*), meningkatkan kejelasan interpretasi hasil, serta mengurangi ambiguitas pelabelan pada komentar publik yang bersifat heterogen dan kontekstual di platform YouTube.

3. Berdasarkan hasil ekstraksi metadata komentar, dataset dalam penelitian ini mencakup komentar YouTube yang dipublikasikan dalam rentang waktu Agustus 2020 hingga Desember 2025.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Temuan dari penelitian ini dapat menjadi opsi bagi pendidik dan institusi pendidikan dalam memahami persepsi publik terhadap penerapan kecerdasan buatan (AI) di bidang pendidikan sebagai dasar evaluasi dan pengambilan keputusan.
2. Penelitian ini dapat dijadikan alternatif bagi pengembang teknologi dan pembuat kebijakan pendidikan dalam merumuskan kebijakan serta strategi pengembangan dan pemanfaatan AI yang adaptif, etis, dan selaras dengan kebutuhan masyarakat.

BAB II
STUDI PUSTAKA

2.1 Klasifikasi Sentimen terhadap AI di Bidang Pendidikan

Penelitian ini didukung oleh berbagai studi terdahulu yang relevan, yang memperkuat fokus pada analisis sentimen persepsi masyarakat di platform YouTube terkait kecerdasan buatan (AI) dalam bidang pendidikan. Berikut ini beberapa penelitian yang menjadi dasar dan acuan studi ini.

Penelitian oleh Jamil *et al*. (2024) berangkat dari permasalahan kurangnya pemahaman mengenai persepsi publik terhadap simulasi kesiapsiagaan bencana yang disebarluaskan melalui kanal YouTube BNPB sebagai sarana edukasi mitigasi bencana. Tujuan penelitian ini adalah mengklasifikasikan komentar publik pada video simulasi bencana menggunakan dua metode klasifikasi teks, yaitu *Naïve Bayes classifier* (NBC) dan *Support Vector Machine* (SVM). *Dataset* dikumpulkan melalui *web crawling* dari 33 video YouTube dengan total 696 komentar, kemudian diseleksi dan diberi label menjadi 204 komentar (112 positif, 43 negatif, dan 49 netral). Data diproses melalui tahapan *cleaning, case folding, tokenizing, stopword removal,* dan *stemming* dengan pembobotan TF-IDF. Hasil pengujian menunjukkan bahwa NBC menghasilkan akurasi tertinggi sebesar 80,4% dengan waktu eksekusi tercepat (0,0097 detik), sedangkan SVM memperoleh akurasi 72,3% dengan waktu eksekusi lebih lama (193,48 detik). Kesimpulannya, NBC lebih efisien dan akurat untuk dataset kecil, sedangkan SVM lebih stabil pada dataset besar.

Selanjutnya, Nahid *et al*. (2024) mengkaji analisis persepsi dan emosi publik pada komentar YouTube dalam konteks Pemilihan Presiden Indonesia 2024, dengan latar belakang meningkatnya peran YouTube sebagai ruang opini publik politik yang sarat dengan ekspresi emosi. Penelitian ini bertujuan untuk menganalisis emosi publik secara mendalam melalui pendekatan *multi-label sentiment analysis* terhadap 32.000 komentar YouTube yang dikumpulkan dari kanal KPU RI dan Najwa Shihab, yang diklasifikasikan ke dalam delapan kategori emosi, yaitu *anger, anticipation, disgust, joy, fear, sadness, surprise,* dan *trust*.

Sebelum tahap klasifikasi, data komentar diproses melalui tahapan *pre-processing* teks yang meliputi pembersihan karakter tidak relevan (URL, simbol, dan angka), *case folding*, penghapusan *stopwords*, serta normalisasi teks informal dan singkatan untuk meningkatkan kualitas data. Proses pelabelan sentimen dilakukan secara otomatis menggunakan GPT-3.5 guna menjaga konsistensi klasifikasi emosi, sementara *text augmentation* diterapkan melalui teknik penghapusan dan pertukaran kata secara acak untuk memperkaya variasi data dan meningkatkan kemampuan generalisasi model. Penelitian ini menguji tiga model klasifikasi, yaitu CNN, Bi-LSTM, dan CNN–BiLSTM, dengan hasil terbaik diperoleh oleh Bi-LSTM yang mencapai akurasi 98,4% dan AUC 0,91. Hasil penelitian menunjukkan bahwa model *deep learning multi-label* efektif dalam memahami opini publik terhadap isu politik, serta berpotensi dikembangkan lebih lanjut menggunakan pendekatan *transformer* seperti BERT.

Selain itu, Xu *et al*. (2025) meneliti persepsi publik terhadap teknologi *deepfake* melalui analisis posting dan komentar media sosial Reddit, yang dilatarbelakangi oleh meningkatnya kontroversi etika dan hukum terkait penyalahgunaan *deepfake*. Dataset yang digunakan berjumlah 17.720 postingan dan komentar yang dianalisis menggunakan pendekatan *topic modelling* berbasis *Latent Dirichlet Allocation* (LDA) serta analisis sentimen menggunakan *TextBlob* dan *VADER*. Sebelum tahap analisis, data teks melalui proses *pre-processing* standar untuk meningkatkan kualitas dan konsistensi data. Tahapan tersebut meliputi pembersihan karakter tidak relevan seperti URL, simbol, angka, dan tanda baca menggunakan ekspresi reguler, konversi seluruh teks ke huruf kecil (*case folding*), serta tokenisasi berbasis spasi. Selanjutnya, kata-kata penghenti (*stopwords*) dihapus menggunakan daftar *stopwords* dari pustaka NLTK, diikuti dengan penghapusan kata-kata pendek yang tidak informatif. Proses lemmatisasi juga diterapkan untuk mengubah kata ke bentuk dasarnya guna menyatukan variasi morfologis, serta dilakukan pemeriksaan dan penghapusan data duplikat. Data hasil *pre-processing* kemudian disimpan dalam format terstruktur untuk keperluan analisis lanjutan. Penelitian ini menguji enam algoritma *machine learning* (SVM, *Naïve Bayes*, *Random Forest*, *Decision Tree*, *Logistic*

Regression, dan KNN) serta tiga model berbasis BERT, dengan hasil terbaik diperoleh oleh BERTweet yang mencapai akurasi sebesar 87,03%. Hasil analisis menunjukkan bahwa persepsi publik terhadap *deepfake* terbagi antara pandangan positif terkait kreativitas teknologi dan kekhawatiran terhadap potensi penyalahgunaannya, sehingga penelitian lanjutan diarahkan pada aspek regulasi dan edukasi etis penggunaan kecerdasan buatan.

Kemudian, Alawi & Bozkurt (2024) memberikan kontribusi berbeda dengan menganalisis sentimen publik terhadap universitas di Turki melalui platform Twitter dengan dataset berjumlah 17.793 tweet (13.219 positif dan 4.574 negatif) yang dianotasi secara manual. Karena tweet bersifat informal dan berisik, data di*pre-processing* melalui pembersihan teks (menghapus nama pengguna, URL, karakter khusus, angka, serta *tweet* non-Turki), normalisasi teks (huruf kecil dan huruf berulang), tokenisasi dan penghapusan stopwords, lematisasi menggunakan *snowballstemmer*, serta penyaringan manual untuk menghilangkan kebisingan. Setelah data bersih, fitur diekstraksi menggunakan TF-IDF dan Bag of Words (BoW). Penelitian membandingkan sembilan algoritma machine learning konvensional dan beberapa model deep learning, dengan model hibrida BERT–BiLSTM–CNN menunjukkan performa terbaik (akurasi 0,9101, F1-score 0,8801, AUC-ROC 0,9632), menegaskan efektivitas model hibrida dalam menangkap konteks linguistik yang kompleks.

Sementara itu, Bello *et al*. (2023) berangkat dari keterbatasan pendekatan tradisional seperti Word2Vec, CNN, dan RNN dalam memahami konteks kata dua arah, dan mengembangkan sistem klasifikasi sentimen berbasis BERT untuk mengidentifikasi tiga kategori sentimen: positif, netral, dan negatif. Dalam penelitian ini, dataset terdiri dari 213.355 tweet yang dikumpulkan dari enam dataset yang digabung dari Kaggle, kemudian dilakukan pemrosesan awal seperti eliminasi nilai kosong dan pemetaan label sentimen, serta pre-processing teks dengan menghapus karakter khusus, tanda baca, angka, simbol, dan hashtag agar tidak mengganggu representasi model. Selanjutnya teks ditokenisasi dan setiap token diubah menjadi representasi vektor melalui BERT sehingga konteks kata dapat ditangkap dua arah dengan lebih baik dibandingkan model lain. Hasil

penelitian menunjukkan bahwa kombinasi BERT–BiLSTM memberikan hasil terbaik dengan akurasi sekitar 93% dan F1-score sekitar 95%, jauh melampaui pendekatan dengan Word2Vec yang hanya mencapai 48–57%. Temuan ini menegaskan keunggulan BERT dalam menangkap konteks semantik yang lebih dalam dalam analisis sentimen tweet, dan peneliti menyarankan penerapan model ini untuk bahasa lain guna meningkatkan generalisasi performa model.

Di sisi lain, Straton *et al*. (2024) menyoroti peran media sosial dalam membentuk opini publik terkait isu kesehatan dan stigma sosial selama pandemi COVID-19, khususnya dalam kaitannya dengan *engagement* pengguna terhadap sentimen yang mengandung stigma. Penelitian ini menggunakan jutaan interaksi data dari berbagai platform media sosial seperti Facebook (Meta), Twitter (X), YouTube, dan Reddit untuk mengevaluasi hubungan antara gaya bahasa, fitur linguistik, dan respons publik terhadap konten kesehatan. Sebelum dilakukan analisis kuantitatif, data teks melalui tahapan *pre-processing* yang meliputi pembersihan teks dari elemen-elemen yang tidak relevan seperti URL, karakter khusus, angka, dan penyebutan akun, serta konversi teks ke huruf kecil untuk konsistensi analisis. Selain itu, dilakukan penghapusan konten dalam bahasa yang tidak diinginkan untuk memastikan fokus pada teks utama, dan pembersihan kata-kata umum (*stopwords*) guna mengurangi kebisingan data dalam pemodelan selanjutnya. Setelah data bersih, analisis bahasa dan fitur teks dilakukan menggunakan *Linguistic Inquiry and Word Count* (LIWC) untuk mengekstraksi ciri linguistik yang relevan, yang kemudian dianalisis lebih lanjut menggunakan teknik statistik seperti *ANOVA F-score regression*, *Z-score analysis*, dan pengelompokan *K-Means clustering* untuk mengidentifikasi pola keterlibatan pengguna terhadap sentimen stigma. Hasil penelitian menunjukkan bahwa sebelum pandemi, sentimen negatif cenderung meningkatkan *engagement*, tetapi tren ini menurun selama pandemi, sehingga gaya bahasa dan konteks sosial berpengaruh signifikan terhadap respons publik. Penelitian ini memberikan wawasan penting tentang bagaimana sentimen yang mengandung stigma memengaruhi keterlibatan di berbagai platform sosial dan membuka arahan untuk memperluas dataset lintas bahasa dalam studi lanjutan.

Lebih jauh, Sebők *et al*. (2024) meneliti pengaruh geopolitik terhadap pemberitaan vaksin COVID-19 di media Hungaria dengan menganalisis 72.339 artikel berita dari tiga portal berita utama pada periode Maret 2020 hingga Maret 2022 menggunakan pendekatan *AI-supported sentiment analysis*. Penelitian ini bertujuan untuk menguji apakah orientasi politik media memengaruhi representasi vaksin dari blok geopolitik Timur dan Barat dalam pemberitaan, yang selanjutnya berpotensi membentuk opini publik. Sebelum analisis sentimen dilakukan, data melalui tahapan *pre-processing* teks untuk memastikan kualitas input, meliputi penghapusan elemen non-teks seperti simbol, angka, dan karakter khusus, penghapusan tautan, serta konversi teks ke huruf kecil agar pemrosesan model berlangsung secara konsisten. Data yang telah dibersihkan kemudian diklasifikasikan ke dalam tiga label sentimen, yaitu positif, negatif, dan netral, menggunakan model emBERT yang telah di-*fine-tune* dengan data berlabel manual berbahasa Hungaria. Evaluasi performa model menunjukkan bahwa emBERT mencapai akurasi sebesar 74% dan nilai F1-score sebesar 73%**,** yang mencerminkan kinerja klasifikasi yang stabil dan mendekati standar riset dalam konteks *computational social science*. Hasil penelitian mengungkap adanya variasi sentimen antar portal berita terhadap vaksin dari kelompok geopolitik yang berbeda, yang mengindikasikan pengaruh orientasi politik media dalam membingkai isu kesehatan publik. Temuan ini tidak hanya memberikan pemahaman empiris mengenai dinamika pemberitaan vaksin dalam konteks geopolitik, tetapi juga menyediakan dataset serta model terlatih yang berpotensi dimanfaatkan untuk penelitian lanjutan pada bidang analisis sentimen media dan politik kesehatan.

Selanjutnya, Hirata & Matsuda (2023) berangkat dari krisis tenaga kerja dan gangguan rantai pasok logistik di Jepang akibat pandemi COVID-19 untuk menganalisis persepsi publik terhadap isu logistik pascapandemi melalui data *Twitter*. Penelitian ini menggunakan 17.597 *tweet* berbahasa Jepang yang diambil berdasarkan kata kunci terkait logistik, dengan penghapusan *retweet* dan *tweet* tidak valid sebelum analisis, sehingga data bersih siap diproses. Karena data *Twitter* bersifat tidak terstruktur dan informal, dilakukan serangkaian tahapan

pre-processing teks, antara lain pembersihan data dengan menghapus URL, karakter khusus, angka, dan penyebutan akun, serta konversi teks ke huruf kecil; tokenisasi menggunakan alat analisis morfologi bahasa Jepang seperti MeCab untuk memecah kalimat menjadi unit kata yang bermakna; serta penghapusan kata-kata umum (*stopwords*) yang kurang memiliki bobot semantik dalam konteks analisis. Dengan data yang telah diproses ini, penelitian menerapkan model BERT bahasa Jepang (Daigo BERT *base Japanese-sentiment*) untuk mengkategorikan sentimen, dan hasilnya menunjukkan dominasi sentimen positif terhadap kebijakan pemerintah terkait *"White Logistics".* Temuan ini menegaskan bahwa media sosial efektif digunakan untuk memantau opini publik secara *real time* dalam konteks isu logistik pascapandemi di Jepang, sekaligus membuka arah untuk studi lanjutan dengan pendekatan linguistik dan model yang lebih adaptif terhadap variasi bahasa.

Kemudian, Nayak *et al*. (2023) mengusulkan integrasi *Bayesian Boosting* dengan *Weight-Guided Optimal Feature Selection* untuk mengatasi rendahnya akurasi model tradisional dalam analisis sentimen terhadap ulasan hotel. Penelitian ini menggunakan dataset 515.000 ulasan hotel di Eropa dari *Kaggle* yang diklasifikasi menjadi tiga kategori sentimen: positif, negatif, dan netral, dengan tujuan meningkatkan akurasi klasifikasi dibandingkan model seperti Random Forest dan C4.5. Sebelum dilakukan pemodelan, data teks melalui tahapan *pre-processing* yang mencakup pembersihan data seperti penghapusan nilai kosong dan karakter non-teks (simbol, angka, dan tanda baca), normalisasi teks dengan konversi ke huruf kecil dan penghapusan *stopwords,* serta tokenisasi kata untuk menyederhanakan representasi teks sebagai unit kata yang bermakna. Selanjutnya dilakukan ekstraksi fitur untuk memilih atribut yang paling informatif berdasarkan bobotnya sehingga model dapat mempelajari pola sentimen secara efektif. Dengan pendekatan ini, model *modified Bayesian Boosting* yang dipadukan dengan *weight-guided feature selection* menunjukkan akurasi tinggi sebesar 98,49%**,** jauh melampaui Random Forest (78,6%) dan C4.5 (68,58%). Hasil penelitian ini menunjukkan bahwa teknik optimasi fitur dengan Bayesian

Boosting efektif dalam menangkap pola sentimen pada dataset besar dan dapat menjadi rujukan untuk pengembangan metode serupa dalam analisis opini publik.

Terakhir, Atak *et al*. (2023) meneliti pengaruh sentimen laporan keuangan terhadap nilai perusahaan di Borsa Istanbul (BIST) periode 1998–2022, dengan tujuan menilai hubungan antara sentimen, kinerja keuangan, dan nilai perusahaan. Dataset mencakup teks laporan tahunan dari 85 perusahaan berbahasa Inggris, dianalisis menggunakan tiga model *deep learning* berbasis *transformer*, yaitu FinBERT, DistilFinRoBERTa, dan FinRoBERTa. Sebelum analisis, dilakukan pra-pemrosesan data yang meliputi pembersihan teks dari angka, simbol, dan karakter khusus; penghapusan *stopwords*; normalisasi teks; tokenisasi; serta konversi kata ke bentuk dasarnya (*lemmatization*) untuk menjaga konsistensi dan konteks semantik. Ekstraksi fitur dilakukan melalui representasi *embedding* dari masing-masing model *transformer*. Hasil penelitian menunjukkan FinRoBERTa memberikan performa terbaik dengan akurasi 97,5% dan F1-score 97,7%. Analisis sentimen mengungkapkan bahwa sentimen positif berkontribusi pada peningkatan nilai perusahaan (Tobin's Q dan ROE), sedangkan sentimen negatif menurunkannya. Penelitian ini menegaskan efektivitas analisis sentimen berbasis *deep learning* dalam menilai persepsi pasar dan mengurangi asimetri informasi.

Berbeda dengan penelitian-penelitian tersebut, penelitian ini menitikberatkan pada klasifikasi sentimen komentar publik di YouTube mengenai kecerdasan buatan (AI) dalam konteks pendidikan dengan memanfaatkan IndoBERTa dan BERT, yang memungkinkan representasi semantik dua arah pada teks berbahasa Indonesia, sekaligus menangani informalitas dan keunikan komentar di platform video, sehingga fokusnya lebih spesifik pada persepsi publik terhadap AI dalam pendidikan dibandingkan studi terdahulu yang beragam baik dari topik, bahasa, maupun platform.

2.2 Perkembangan Model Transformer

Perkembangan model dalam pemrosesan bahasa alami *(Natural Language Processing*) menunjukkan adanya pergeseran paradigma yang signifikan, khususnya dari pendekatan berbasis jaringan syaraf tiruan konvensional menuju

arsitektur yang lebih mampu menangkap kompleksitas konteks bahasa secara global (Sun *et al*., 2022). Pada tahap awal, model seperti *Artificial Neural Network* (ANN) digunakan untuk mengenali pola dasar dalam data teks, namun belum mampu menangani sifat sekuensial bahasa secara optimal (Schoene *et al*., 2022). Selanjutnya, *Recurrent Neural Network* (RNN) dan *Long Short-Term Memory* (LSTM) dikembangkan untuk mengatasi permasalahan tersebut dengan memanfaatkan ketergantungan temporal antar token dalam suatu urutan teks. Meskipun demikian, kedua model ini masih memiliki keterbatasan, terutama dalam menangkap dependensi jangka panjang, menghadapi permasalahan *vanishing gradient*, serta membutuhkan waktu pelatihan yang relatif lama ketika diterapkan pada data berskala besar dan kompleks (Ghosh *et al*., 2025).

Seiring dengan meningkatnya kebutuhan akan model yang lebih efisien dan mampu memahami konteks secara mendalam, arsitektur *Transformer* diperkenalkan sebagai solusi inovatif dalam NLP modern. *Transformer* mengandalkan mekanisme *self-attention* yang memungkinkan model untuk mengevaluasi hubungan antar token secara simultan dalam satu waktu, tanpa bergantung pada urutan sekuensial. Hal ini memungkinkan model untuk menangkap hubungan semantik dan sintaksis secara global, bahkan pada teks yang panjang dan kompleks. Selain itu, kemampuan pemrosesan paralel yang dimiliki *Transformer* secara signifikan meningkatkan efisiensi komputasi, sehingga proses pelatihan menjadi lebih cepat dan skalabel. Berbagai penelitian menunjukkan bahwa arsitektur ini secara konsisten mampu mengungguli model sekuensial tradisional dalam berbagai tugas NLP, seperti klasifikasi teks, *part-of-speech tagging*, dan ekstraksi informasi kompleks (Ghosh *et al*., 2025; Sari & Abidin, 2026).

Arsitektur *Transformer* kemudian menjadi fondasi utama bagi pengembangan model bahasa pralatih *(pre-trained language models)*, seperti BERT *(Bidirectional Encoder Representations from Transformers)*, serta berbagai variannya yang diadaptasi untuk bahasa tertentu, termasuk IndoRoBERTa dan IndoBERT untuk bahasa Indonesia. Dalam konteks bahasa Indonesia, pemanfaatan model berbasis *Transformer* menunjukkan peningkatan performa

yang signifikan dibandingkan pendekatan sebelumnya, terutama dalam tugas seperti klasifikasi teks dan pengenalan entitas bernama. Hal ini disebabkan oleh kemampuan model dalam menghasilkan representasi kontekstual yang lebih kaya melalui proses *pre-training* pada korpus besar. Secara empiris, model *Transformer* terbukti memiliki kemampuan generalisasi yang lebih baik dalam berbagai aplikasi NLP modern (Yulianti *et al*., 2024;. Putri & Ardiansyah, 2023).

Selain itu, studi terkini menunjukkan bahwa model berbasis *Transformer* seperti BERT dan RoBERTa tidak hanya unggul dalam tugas klasifikasi, tetapi juga efektif dalam berbagai domain spesifik seperti analisis naratif dan prediksi sentimen berbasis konteks, yang semakin memperkuat posisinya sebagai pendekatan dominan dalam NLP modern (Poleksić & Martinčić-ipšić, 2025).

2.2.1 Model Transformer

Model Transformer merupakan arsitektur jaringan syaraf tiruan yang dirancang untuk memproses data sekuensial tanpa menggunakan mekanisme rekurensi seperti pada RNN atau LSTM (Mienye *et al*., 2024). Pendekatan utama yang digunakan adalah mekanisme *attention*, khususnya *self-attention*, yang memungkinkan model untuk memahami hubungan antar token dalam suatu teks secara langsung dan menyeluruh (Rahman et *al*., 2024). Berbeda dengan model sekuensial yang memproses token satu per satu berdasarkan urutan waktu, Transformer memproses seluruh token secara paralel, sehingga tidak hanya meningkatkan efisiensi komputasi tetapi juga memungkinkan model untuk menangkap keterkaitan antar kata secara simultan dalam satu representasi (Wang *et al*., 2025). Kemampuan ini menjadi krusial dalam pemrosesan bahasa alami yang memiliki struktur kompleks, karena makna suatu kata sering kali dipengaruhi oleh keseluruhan konteks kalimat, bukan hanya oleh kata-kata di sekitarnya secara lokal (Bettayeb *et al*., 2024; Zi et al., 2025; Pfeffer *et al*., 2025).

Kemampuan utama Transformer terletak pada mekanisme *self-attention* yang memungkinkan model memahami hubungan antar token secara global dalam satu waktu (Tucudean *et al*., 2024). Berbeda dengan RNN dan LSTM yang memproses data secara sekuensial, Transformer mampu melakukan pemrosesan

paralel sehingga lebih efisien dalam menangani dataset berskala besar dan teks panjang (Zhang & Shafiq, 2024). Pendekatan ini menjadi fondasi utama dalam pengembangan model *pretrained* modern seperti BERT dan RoBERTa yang saat ini banyak digunakan pada berbagai tugas *Natural Language Processing* (Gardazi *et al*., 2025). termasuk klasifikasi sentimen, *question answering*, dan *text generation* (Tucudean *et al*., 2024).

Secara arsitektural, Transformer terdiri atas beberapa komponen utama, yaitu lapisan *embedding*, mekanisme *multi-head self-attention*, serta jaringan *feed-forward*. Lapisan *embedding* berfungsi untuk mengubah token menjadi representasi numerik dalam ruang vektor berdimensi tinggi, yang kemudian diproses lebih lanjut oleh mekanisme *attention* (Li *et al*., 2025). *Multi-head self-attention* memungkinkan model untuk mengeksplorasi berbagai hubungan kontekstual dalam beberapa subruang representasi secara bersamaan, sehingga model dapat menangkap pola linguistik yang beragam, seperti hubungan sintaksis, semantik, maupun dependensi jarak jauh (Adablanu *et al*., 2025). Sementara itu, jaringan *feed-forward* berperan dalam melakukan transformasi non-linear terhadap hasil *attention*, sehingga menghasilkan representasi yang lebih abstrak dan informatif. Kombinasi dari komponen-komponen ini menjadikan *Transformer* mampu membangun pemahaman bahasa yang lebih komprehensif dibandingkan pendekatan sebelumnya (Tay *et al*., 2022; Zhang & Shafiq, 2024; Chen *et al*., 2024).

Selain itu, *Transformer* juga memanfaatkan *positional encoding* untuk mempertahankan informasi urutan token, yang menjadi aspek penting dalam pemahaman bahasa alami. Hal ini diperlukan karena mekanisme *self-attention* pada dasarnya tidak memiliki kesadaran terhadap urutan posisi token. Dengan menambahkan informasi posisi ke dalam representasi *embedding*, model tetap mampu memahami struktur urutan kata dalam kalimat, sehingga interpretasi makna tidak kehilangan konteks sintaksisnya. Pendekatan ini memungkinkan *Transformer* untuk menggabungkan keunggulan pemrosesan paralel dengan pemahaman urutan bahasa secara efektif (He *et al*., 2021; Chu *et al*., 2023).

Keunggulan tersebut menjadikan *Transformer* sangat efektif dalam menangkap dependensi jangka panjang tanpa mengalami degradasi performa seperti yang sering terjadi pada model berbasis RNN atau LSTM. Selain itu, kemampuan pemrosesan paralel membuat model ini lebih efisien dan *scalable* ketika diterapkan pada dataset berskala besar. Oleh karena itu, *Transformer* menjadi fondasi utama dalam pengembangan model NLP modern dan secara konsisten menunjukkan performa yang lebih baik dalam berbagai tugas, baik dari segi akurasi maupun efisiensi komputasi (Tucudean *et al*., 2024; Sari & Abidin, 2026).

2.2.2 BERT

BERT (*Bidirectional Encoder Representations from Transformers*) merupakan model bahasa pralatih yang dikembangkan berdasarkan arsitektur *Transformer encoder* (Siwinska *et al.,* 2026). Model ini dirancang untuk menghasilkan representasi bahasa yang bersifat dua arah (*bidirectional*), di mana setiap token dipahami dengan mempertimbangkan konteks sebelum dan sesudahnya secara simultan (Lee *et al*., 2022). Pendekatan ini memungkinkan BERT untuk menangkap makna kata secara lebih akurat dan kontekstual, terutama dalam kasus ambiguitas leksikal, di mana satu kata dapat memiliki makna yang berbeda tergantung pada konteks penggunaannya dalam kalimat (Poleksić & Martinčić-ipšić, 2025).

Keunggulan utama BERT dibandingkan model NLP sebelumnya terletak pada kemampuannya memahami konteks kata secara dua arah (*bidirectional context*) (Gardazi *et al.,* 2025). Pendekatan ini memungkinkan setiap token dipahami berdasarkan hubungan dengan token sebelum dan sesudahnya secara simultan, sehingga representasi semantik yang dihasilkan menjadi lebih kontekstual dan akurat (Subakti *et al*., 2022). Selain itu, BERT menggunakan pendekatan *Masked Language Modeling* (MLM) pada tahap *pre-training*, yaitu dengan menyembunyikan sebagian token dalam kalimat dan meminta model memprediksi token yang hilang berdasarkan konteks kalimat (Higuchi *et al*., 2022; Gardazi *et al.,* 2025). Strategi tersebut memungkinkan model mempelajari

struktur bahasa, hubungan sintaktis, serta keterkaitan semantik secara lebih mendalam dibandingkan pendekatan *sequential* tradisional yang hanya memproses teks dalam satu arah (Subakti *et al*., 2022; Islam et al., 2026). Melalui kombinasi mekanisme *bidirectional encoding* dan MLM, BERT mampu menghasilkan *contextual word representations* yang lebih efektif untuk berbagai tugas NLP, termasuk klasifikasi teks, analisis sentimen, *question answering*, dan *named entity recognition* (Gardazi *et al*., 2025; Poleksić & Martinčić-ipšić, 2025).

Pendekatan *bidirectional* pada BERT menjadi inovasi penting dalam NLP karena mampu mengatasi keterbatasan model sebelumnya yang hanya memahami konteks secara satu arah. Dengan mempertimbangkan seluruh konteks kalimat secara bersamaan, BERT mampu membangun representasi semantik yang lebih kaya, mendalam, dan sensitif terhadap hubungan antar kata. Hal ini sangat penting dalam tugas-tugas yang membutuhkan pemahaman bahasa tingkat tinggi, seperti analisis sentimen berbasis konteks, pemahaman pertanyaan, serta ekstraksi informasi kompleks (Subakti *et al*., 2022; Gardazi *et al*., 2025).

Dalam proses *pre-training*, BERT menggunakan strategi seperti *masked language modeling* dan *next sentence prediction*, yang memungkinkan model mempelajari hubungan intra-kalimat dan antar kalimat secara efektif. Pendekatan ini tidak hanya meningkatkan kemampuan model dalam memahami struktur bahasa, tetapi juga memperkuat generalisasi model ketika diterapkan pada berbagai tugas spesifik melalui proses *fine-tuning*. Dengan demikian, BERT tidak hanya berfungsi sebagai model tunggal, tetapi juga sebagai fondasi yang dapat disesuaikan untuk berbagai aplikasi NLP (Subakti *et al.*, 2022).

Penelitian terbaru menunjukkan bahwa BERT tetap menjadi salah satu model paling berpengaruh dalam NLP, dengan berbagai pengembangan lanjutan yang meningkatkan efisiensi, ukuran model, serta kemampuan adaptasi terhadap domain tertentu. Model ini juga telah digunakan dalam berbagai bidang, termasuk analisis data ilmiah dan domain spesifik lainnya yang membutuhkan pemahaman bahasa berbasis konteks (Poleksić & Martinčić-ipšić, 2025). Selain itu, dalam berbagai eksperimen komparatif, BERT secara konsisten menunjukkan performa tinggi dalam tugas klasifikasi teks dan pemahaman bahasa, sehingga tetap relevan

sebagai baseline maupun model utama dalam penelitian NLP modern *(Gardazi et al*., 2025).

2.2.3 IndoRoBERTa

IndoRoBERTa merupakan pengembangan dari model RoBERTa yang secara khusus diadaptasi untuk bahasa Indonesia dengan tetap mempertahankan arsitektur Transformer *encoder* (Vincent *et al.*, 2025). Model ini dikembangkan sebagai respons terhadap keterbatasan model global yang umumnya dilatih menggunakan korpus berbahasa Inggris, sehingga kurang mampu menangkap karakteristik linguistik bahasa Indonesia secara optimal. Dengan adanya IndoRoBERTa, representasi bahasa yang dihasilkan menjadi lebih sesuai dengan struktur, pola, dan variasi bahasa Indonesia yang digunakan dalam konteks nyata (Moghe *et al*., 2021).

Dalam proses pelatihannya, IndoRoBERTa menggunakan korpus besar berbahasa Indonesia yang mencakup berbagai jenis teks, baik formal maupun informal. Hal ini memungkinkan model untuk memahami variasi penggunaan bahasa, termasuk slang, singkatan, serta struktur kalimat yang tidak baku yang sering muncul dalam media sosial. Dengan demikian, model tidak hanya mampu memahami bahasa dalam konteks formal, tetapi juga adaptif terhadap dinamika bahasa sehari-hari yang lebih kompleks dan variatif (Hessel *et al*., 2021; Moghe *et al*., 2021).

Berbagai penelitian menunjukkan bahwa IndoRoBERTa memiliki performa yang lebih unggul dibandingkan model deep learning konvensional, terutama dalam tugas analisis sentimen dan klasifikasi teks. Keunggulan ini disebabkan oleh kemampuan model dalam menghasilkan representasi kontekstual yang lebih kaya serta kemampuannya dalam menangkap hubungan semantik secara lebih mendalam (Wicaksono & Rojali, 2026). Studi terbaru juga menunjukkan bahwa RoBERTa dan variannya mampu mengungguli BERT dalam beberapa skenario, seperti klasifikasi multilabel dan penanganan data tidak seimbang, yang menunjukkan kemampuan generalisasi yang lebih baik dalam kondisi data yang kompleks (Sari & Abidin, 2026).

Dengan kemampuan representasi kontekstual yang kuat serta adaptasi terhadap karakteristik bahasa Indonesia, IndoRoBERTa menjadi model yang sangat relevan untuk analisis sentimen teks berbahasa Indonesia, khususnya pada data media sosial seperti komentar YouTube yang cenderung tidak terstruktur. Keunggulan ini menjadikan IndoRoBERTa sebagai salah satu model unggulan dalam NLP berbahasa Indonesia dan banyak digunakan dalam penelitian terkini yang berfokus pada analisis teks berbasis konteks (Moghe *et al*., 2021; *Yulianti et al*., 2024).

2.2.4 Perbedaan BERT dan IndoRoBERTa

Meskipun BERT dan IndoRoBERTa sama-sama menggunakan arsitektur Transformer *encoder*, kedua model memiliki beberapa perbedaan mendasar pada proses *pretraining*, jenis korpus, tokenizer, dan optimasi pelatihan (Tarumingkeng, 2024). BERT menggunakan pendekatan *Masked Language Modeling* (MLM) dan *Next Sentence Prediction* (NSP) dalam proses *pretraining* (Jullfikar *et al*., 2025). Sementara itu, IndoRoBERTa menghilangkan NSP dan menggunakan *dynamic masking* yang memungkinkan proses pelatihan menjadi lebih optimal (M. I. Abidin & Pamungkas, 2025).

Selain itu, BERT pada umumnya dilatih menggunakan korpus multilingual atau bahasa Inggris (El Majid & Nuari, 2025), sedangkan IndoRoBERTa dilatih secara khusus menggunakan korpus besar bahasa Indonesia sehingga lebih mampu memahami karakteristik linguistik lokal dan bahasa informal pada media sosial (Suprianto & Ningsih, 2025). Perbedaan tersebut menyebabkan IndoRoBERTa memiliki kemampuan representasi semantik yang lebih baik untuk teks berbahasa Indonesia dibandingkan model multilingual umum.

Tabel 2.1 berikut menunjukkan perbedaan utama antara BERT dan IndoRoBERTa berdasarkan aspek arsitektur dan proses pretraining.

Tabel 2. 1 Perbedaan Arsitektur BERT dan IndoRoBERTa

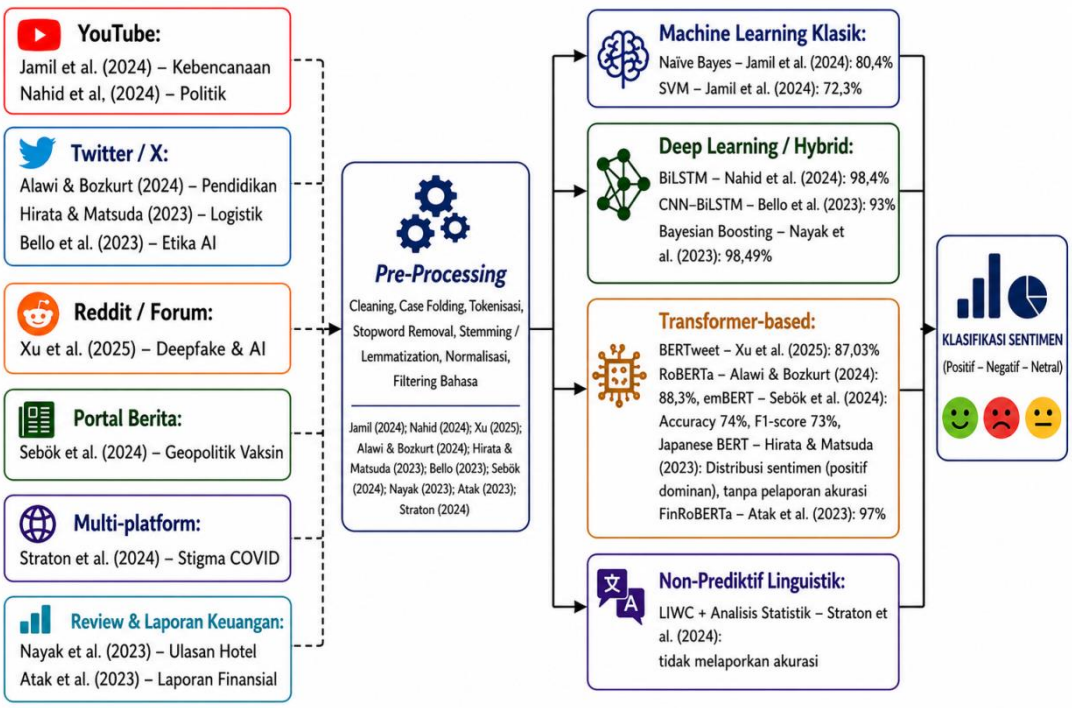
Aspek	BERT	IndoRoBERTa
Arsitektur	Transformer Encoder	Transformer Encoder
Objective Pretraining	MLM + NSP	MLM tanpa NSP
Masking	Static Masking	Dynamic Masking
Tokenizer	WordPiece	Byte Pair Encoding (BPE)
Bahasa Pelatihan	Multilingual/Inggris	Bahasa Indonesia
Corpus	General Corpus	Indonesian Corpus
Segment Embedding	Ya	Tidak digunakan
Adaptasi Bahasa Indonesia	Sedang	Tinggi

Berdasarkan Tabel 2.1, perbedaan utama antara BERT dan IndoRoBERTa tidak terletak pada struktur *Transformer Encoder* yang digunakan, melainkan pada strategi *pretraining*, teknik tokenisasi, mekanisme *masking*, dan jenis korpus pelatihan yang digunakan selama proses pembelajaran model. Karakteristik tersebut menyebabkan IndoRoBERTa lebih adaptif terhadap variasi bahasa Indonesia yang umum ditemukan pada komentar media sosial.

2.3 Kerangka Teori

Kerangka teori dalam penelitian ini disusun berdasarkan sintesis temuan penelitian terdahulu yang mengkaji analisis sentimen menggunakan data teks dari berbagai platform digital. Kerangka ini bertujuan memberikan gambaran konseptual mengenai alur analisis sentimen, mulai dari sumber data, tahapan *pre-processing*, pendekatan metode analisis, hingga keluaran berupa klasifikasi sentimen.

Sebagaimana ditunjukkan pada Gambar 2.1 Kerangka Teori, proses klasifikasi sentimen diawali dari dataset yang bersumber dari beragam platform digital.



Gambar 2. 1 Kerangka Teori

Berdasarkan Gambar 2.1, penelitian terdahulu memanfaatkan komentar pengguna dari YouTube, *Twitter*, forum diskusi seperti Reddit, portal berita daring, hingga dokumen dan ulasan formal. Data dari YouTube banyak digunakan untuk merepresentasikan opini publik terhadap isu kebencanaan dan politik (Jamil *et al*., 2024; Nahid *et al*., 2024), sementara *Twitter* dimanfaatkan untuk analisis pendidikan, isu logistik, etika AI, dan kebijakan publik (Alawi & Bozkurt, 2024; Hirata & Matsuda, 2023; Bello *et al*., 2023). Selain itu, Reddit dan portal berita digunakan untuk menganalisis persepsi publik terhadap *deepfake*, vaksin, dan isu geopolitik (Xu *et al*., 2025; Sebők *et al*., 2024), sedangkan data multi-platform serta dokumen formal digunakan untuk konteks stigma sosial, ulasan hotel, dan laporan keuangan (Straton, 2024; Nayak *et al*., 2023; Atak, 2023).

Tahapan berikutnya adalah *pre-processing*, yang berfungsi meningkatkan kualitas data sebelum dianalisis lebih lanjut. Berdasarkan penelitian terdahulu, tahapan *pre-processing* umumnya meliputi pembersihan teks *(cleaning)* dari URL, simbol, dan karakter non-teks, konversi huruf *(case folding)*, tokenisasi, penghapusan stopword *(stopword removal)*, serta *stemming* atau *lemmatisasi*.

Beberapa penelitian juga menambahkan normalisasi teks informal dan penyaringan bahasa untuk menangani variasi linguistik pada media sosial (Jamil *et al*., 2024; Nahid *et al*., 2024; Xu *et al*., 2025; Bello *et al*., 2023; Hirata & Matsuda, 2023). Tahapan ini menjadi fondasi penting agar representasi teks yang dihasilkan lebih konsisten dan informatif bagi model analisis sentimen.

Setelah data diproses, analisis dilakukan menggunakan beragam pendekatan metode, sebagaimana digambarkan dalam Gambar 2.1. Pendekatan pertama adalah *machine learning* klasik, seperti *Naïve Bayes* dan *Support Vector Machine* (SVM), yang terbukti cukup efektif pada dataset berukuran kecil hingga menengah, dengan akurasi mencapai 80,4% pada Naïve Bayes (Jamil *et al*., 2024). Pendekatan kedua adalah deep learning dan model hibrida, seperti CNN, Bi-LSTM, CNN–BiLSTM, serta *Bayesian Boosting*, yang menunjukkan performa sangat tinggi pada dataset besar dan kompleks, dengan akurasi hingga 98,49% (Nahid *et al.*, 2024; Bello *et al*., 2023; Nayak *et al*., 2023).

Perkembangan terbaru ditandai oleh penggunaan model berbasis *transformer*, seperti BERT dan berbagai variannya (BERTweet, RoBERTa, FinRoBERTa, dan Japanese BERT). Model-model ini mampu menangkap konteks semantik dua arah secara lebih mendalam dan menunjukkan kinerja unggul dalam berbagai domain, dengan akurasi berkisar antara 87% hingga 97,5% (Xu *et al*., 2025; Alawi & Bozkurt, 2024; Atak, 2023). Selain pendekatan prediktif, beberapa penelitian juga menggunakan pendekatan linguistik non-prediktif, seperti LIWC yang dipadukan dengan analisis statistik dan klasterisasi untuk memahami pola bahasa dan keterlibatan pengguna, meskipun tanpa pelaporan akurasi klasifikasi (Straton, 2024).

Seluruh rangkaian proses tersebut bermuara pada output klasifikasi sentimen, yaitu pengelompokan opini publik ke dalam kategori sentimen positif, negatif, dan netral, atau distribusi emosi tertentu sesuai konteks penelitian. Dengan demikian, Gambar 2.1 menegaskan bahwa klasifikasi sentimen merupakan proses terstruktur yang melibatkan integrasi dataset yang relevan, *pre-processing* teks yang tepat, serta pemilihan metode analisis yang sesuai untuk menghasilkan pemahaman yang komprehensif terhadap persepsi publik. Kerangka

teori ini menjadi landasan konseptual bagi penelitian ini dalam menganalisis sentimen komentar publik di YouTube terkait kecerdasan buatan (AI) dalam bidang pendidikan, dengan memanfaatkan dan membandingkan model *transformer* BERT dan IndoRoBERTa guna memperoleh hasil klasifikasi sentimen yang akurat dan kontekstual pada bahasa Indonesia.

Tabel 2.1 berikut menyajikan ringkasan sepuluh penelitian terdahulu yang relevan dengan topik klasifikasi sentimen, mencakup metode, *dataset*, dan temuan utama, yang menjadi dasar serta acuan dalam penelitian ini.

Tabel 2. 2 Ringkasan Penelitian Terdahulu

No	Peneliti (Tahun)	Topik & Dataset	Metode	Hasil Akurasi
1	Jamil *et al.* (2024)	Persepsi publik terhadap simulasi kesiapsiagaan bencana; 696 komentar YouTube BNPB (Indonesia)	Naïve Bayes, SVM (TF-IDF)	NB: 80,4%; SVM: 72,3%
2	Nahid *et al.* (2024)	Emosi dan sentimen publik pada Pilpres Indonesia 2024; ±32.000 komentar YouTube	CNN, BiLSTM, CNN–BiLSTM (multi-label emotion)	BiLSTM: 98,4%
3	Xu *et al.* (2025)	Persepsi publik terhadap teknologi deepfake; 17.720 posting dan komentar Reddit	BERTweet, SVM, NB, RF, DT, LR, KNN	BERTweet: 87,03%
4	Alawi & Bozkurt (2024)	Sentimen publik terhadap universitas di Turki; 17.793 tweet Twitter	BERT–BiLSTM–CNN, ML konvensional	91,01%
5	Bello *et al.* (2023)	Diskursus etika AI di media sosial; 213.355 tweet (Kaggle)	BERT–BiLSTM	±93%
6	Straton *et al.* (2024)	Engagement dan sentimen stigma COVID-19; data lintas platform (Facebook, Twitter, YouTube, Reddit)	LIWC + Statistik (ANOVA, Z-score, K-Means)	Tidak dilaporkan (sentimen negatif berkorelasi kuat dengan tingkat engagement)
7	Sebők *et al.* (2024)	Geopolitik pemberitaan vaksin COVID-19; 72.339 artikel portal berita Hungaria	emBERT (3 kelas sentimen)	74% (F1-score: 73%)
8	Hirata & Matsuda (2023)	Persepsi publik terhadap logistik pascapandemi di Jepang; 17.597 tweet	Japanese BERT (Daigo BERT)	Tidak dilaporkan (dominasi

		Twitter		sentimen positif terhadap kebijakan logistik)
9	Nayak *et al*. (2023)	Analisis sentimen ulasan hotel; 515.000 ulasan Kaggle	Bayesian Boosting + Feature Selection	98,49%
10	Atak *et al.* (2023)	Sentimen laporan keuangan dan nilai perusahaan; laporan tahunan 85 perusahaan	FinBERT, DistilFinRoBERTa, FinRoBERTa	FinRoBERTa: 97,5%

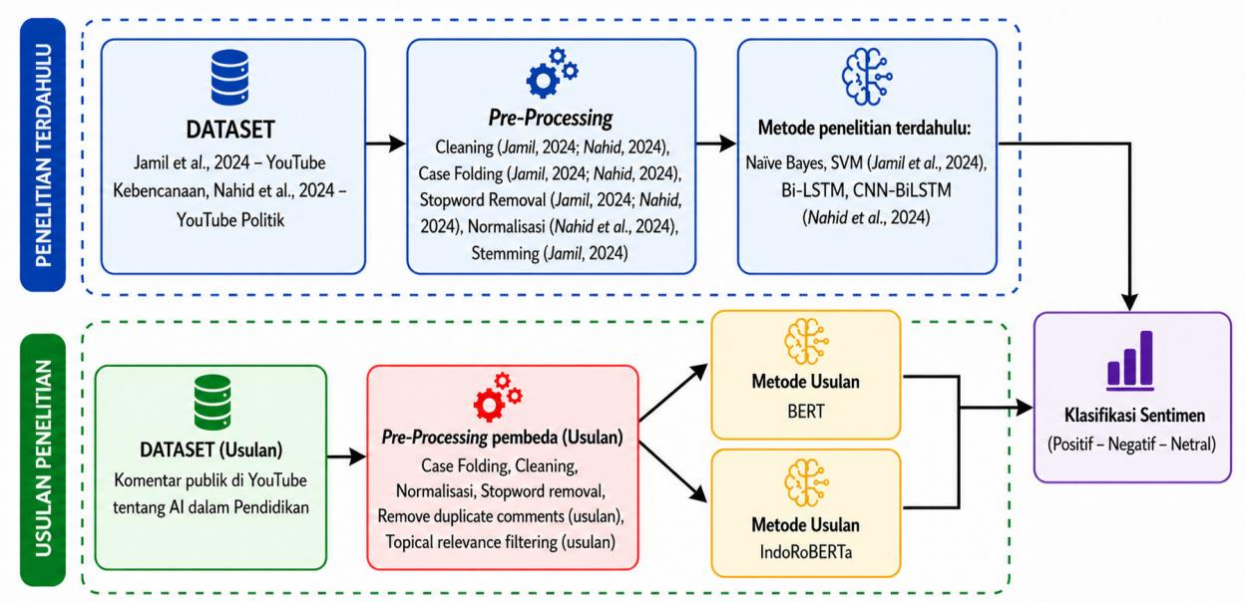
Berdasarkan Tabel 2.2, sebagian besar penelitian terdahulu berfokus pada klasifikasi sentimen komentar publik di media sosial menggunakan berbagai metode, mulai dari *machine learning* tradisional seperti Naïve Bayes dan SVM hingga model deep learning dan Transformer seperti CNN, BiLSTM, BERT, BERTweet, dan FinBERT. Meskipun beberapa penelitian telah menggunakan data komentar YouTube, topik yang dikaji umumnya berkaitan dengan bencana, politik, kesehatan, dan isu sosial lainnya. Penelitian yang secara khusus membandingkan performa BERT dan IndoRoBERTa dalam mengklasifikasikan sentimen komentar publik di YouTube terkait *Artificial Intelligence* (AI) di bidang pendidikan masih sangat terbatas.

Selain itu, berdasarkan kajian pada Tabel 2.2, sebagian besar penelitian belum mengkaji secara mendalam pengaruh karakteristik bahasa Indonesia yang bersifat informal, singkatan, maupun variasi penulisan yang umum ditemukan pada komentar YouTube terhadap kinerja model *pretrained* berbasis Transformer. Oleh karena itu, penelitian ini berfokus pada analisis komparatif antara BERT dan IndoRoBERTa dalam klasifikasi sentimen komentar publik di YouTube terkait AI di bidang pendidikan guna mengetahui model yang memiliki performa terbaik pada konteks bahasa Indonesia.

BAB III
METODOLOGI PENELITIAN

3.1 Kerangka Konsep

Kerangka konsep penelitian ini disusun untuk menggambarkan alur sistematis klasifikasi sentimen komentar publik YouTube terkait kecerdasan buatan (AI) dalam bidang pendidikan, mulai dari tahap pengumpulan data, *pre-processing* teks, penerapan metode klasifikasi sentimen, hingga menghasilkan klasifikasi sentimen akhir, sebagaimana ditunjukkan pada Gambar 3.1**.**



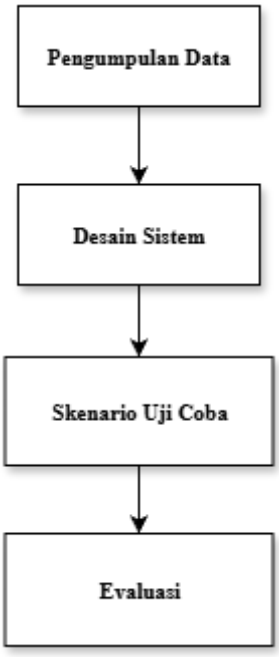
Gambar 3. 1 Kerangka Konsep

Berdasarkan Gambar 3.1, penelitian terdahulu umumnya menggunakan dataset komentar YouTube pada domain kebencanaan dan politik, dengan tahapan *pre-processing* standar meliputi *cleaning*, *case folding*, *stopword removal*, normalisasi, dan *stemming*, serta menerapkan metode klasifikasi seperti *Naïve Bayes*, *Support Vector Machine* (SVM), Bi-LSTM, dan CNN-BiLSTM. Berangkat dari pendekatan tersebut, penelitian ini mengusulkan pengembangan kerangka konsep dengan menggunakan dataset spesifik berupa komentar publik YouTube mengenai kecerdasan buatan (AI) dalam bidang pendidikan. Tahapan *pre-processing* yang diterapkan diperluas melalui penambahan penghapusan komentar duplikat dan *topical relevance filtering* untuk memastikan kesesuaian

konteks data dengan fokus penelitian. Selanjutnya, proses klasifikasi sentimen dilakukan menggunakan model transformer, yaitu BERT dan IndoRoBERTa, untuk mengelompokkan sentimen komentar ke dalam tiga kategori, yaitu positif, negatif, dan netral, sehingga memungkinkan pengukuran kontribusi performa kedua model secara terukur dan relevan dengan konteks analisis sentimen komentar YouTube berbahasa Indonesia.

3.2 Prosedur Penelitian

Prosedur penelitian ini disusun untuk menggambarkan langkah-langkah sistematis yang dilakukan sejak pengumpulan data hingga evaluasi model. Berikut desain prosedur penelitiaan ini dapat dilihat pada gambar 3.2.:



Gambar 3. 2 Prosedur Penelitian

Berdasarkan Gambar 3.2, prosedur penelitian dimulai dari pengumpulan data, dilanjutkan dengan desain sistem, kemudian skenario uji coba, dan terakhir dilakukan evaluasi untuk menilai kinerja model yang digunakan.

3.2.1 Pengumpulan Data

Penelitian ini berfokus pada analisis sentimen komentar publik di platform YouTube yang membahas kecerdasan buatan (AI) dalam bidang pendidikan. YouTube dipilih sebagai sumber data karena merupakan salah satu platform media sosial terbesar yang menyediakan ruang diskusi publik yang aktif dan terbuka, sehingga memungkinkan peneliti mengamati opini, persepsi, serta respons masyarakat terhadap isu-isu teknologi secara luas dan beragam, termasuk pemanfaatan dan dampak AI dalam konteks pendidikan (Jamil *et al*., 2024; Alawi & Bozkurt, 2024).

Pengumpulan data dirancang menggunakan pendekatan *web scraping* komentar YouTube berbasis pencarian kata kunci tematik dengan memanfaatkan YouTube Data API v3 yang diakses melalui skrip Python. Berbeda dengan pendekatan yang berfokus pada pemilihan video tertentu, penelitian ini menggunakan kata kunci sebagai dasar penelusuran untuk menjaring video dan komentar yang relevan dengan topik AI dalam pendidikan. Kata kunci yang digunakan meliputi: *"AI pendidikan", "kecerdasan buatan pendidikan", "AI dalam pendidikan", "AI pembelajaran", "AI di sekolah", "AI di kampus", "AI di dunia pendidikan", "ChatGPT pendidikan", "AI untuk pembelajaran", "dampak AI pada pendidikan", dan "etika AI dalam pendidikan".* Pendekatan ini dirancang untuk memperoleh cakupan data yang lebih luas dan representatif terhadap diskursus publik mengenai AI di ranah pendidikan.

Melalui kata kunci tersebut, sistem secara otomatis menelusuri video YouTube yang relevan dalam rentang waktu publikasi tertentu, kemudian mengekstraksi komentar pengguna dari setiap video yang ditemukan. Setiap komentar direpresentasikan sebagai satu unit data yang mencakup teks komentar serta metadata pendukung, seperti waktu unggah komentar (*publishedAt*), dan identitas video (*videoId*). Seluruh data hasil *scraping* disimpan dalam format XLS sebagai data mentah sebelum memasuki tahapan *pre-proseccing*, penyaringan relevansi, pemodelan berbasis IndoRoBERTa dan BERT, serta evaluasi hasil klasifikasi sentimen.

Berdasarkan hasil ekstraksi metadata komentar, dataset yang dirancang dalam penelitian ini mencakup komentar YouTube yang dipublikasikan dalam rentang waktu Agustus 2020 hingga Desember 2025. Rentang waktu ini memungkinkan analisis persepsi publik yang lebih komprehensif terhadap perkembangan dan pemanfaatan kecerdasan buatan dalam bidang pendidikan, baik sebelum, selama, maupun setelah periode percepatan adopsi teknologi AI di ruang publik digital. Rentang waktu tersebut juga selaras dengan temuan penelitian terkini yang menunjukkan bahwa periode pasca-2020 merupakan fase akselerasi transformasi digital dan adopsi kecerdasan buatan dalam pendidikan secara global, terutama sejak disrupsi akibat pandemi COVID-19, yang secara signifikan meningkatkan interaksi publik terhadap teknologi AI di platform digital (Bond *et al*., 2024; Sevilla *et al*., 2025). Sebagai ilustrasi struktur data yang digunakan, contoh dataset hasil pengumpulan data ditampilkan pada Tabel 3.1.

Tabel 3. 1 Contoh Dataset Komentar YouTube

Comment	Published At	Video ID	keyword	Title	Channel
Hari ini AI masih bayi, kita tunggu si AI beranjak remaja 2030 akan sedahsyat apa dia...	2025-04-07	Glvgjr-gPUQ	AI pendidikan	[FULL] KICK ANDY - PROF STELLA: OTAK VS AI	METRO TV
Sadarkah anda? AI dibuat oleh penguasa dunia agar manusia semakin bergantung...	2025-04-06	Glvgjr-gPUQ	AI pendidikan	[FULL] KICK ANDY - PROF STELLA: OTAK VS AI	METRO TV
Indonesia butuh literasi digital dan AI sejak TK hingga mahasiswa	2025-04-05	Glvgjr-gPUQ	AI pendidikan	[FULL] KICK ANDY - PROF STELLA: OTAK VS AI	METRO TV
Secanggih apapun robot, belum bisa secanggih manusia yang bisa punya anak	2025-03-20	Glvgjr-gPUQ	AI pendidikan	[FULL] KICK ANDY - PROF STELLA: OTAK VS AI	METRO TV
Semoga makin banyak orang seperti Bu Stella...	2025-03-20	Glvgjr-gPUQ	AI pendidikan	[FULL] KICK ANDY - PROF STELLA: OTAK VS AI	METRO TV
…	…	…	…	…	…

Berdasarkan Tabel 3.1, dataset yang digunakan dalam penelitian ini memiliki struktur terorganisir yang terdiri dari beberapa atribut utama, yaitu Comment sebagai teks opini pengguna yang menjadi objek analisis sentimen, Published At sebagai waktu publikasi komentar, *Video ID* sebagai identitas unik video, *keyword* sebagai kata kunci pengambilan data untuk menjaga relevansi topik, serta *Title* dan *Channel* yang memberikan konteks sumber konten. Struktur ini memungkinkan proses analisis dilakukan secara sistematis dengan mempertimbangkan isi komentar sekaligus metadata pendukungnya.

3.2.1.1 Metadata Dataset

Metadata merupakan informasi pendukung yang menyertai setiap komentar YouTube hasil proses crawling menggunakan YouTube Data API v3 (Putra *et al.*, 2025). Metadata digunakan untuk mendukung validitas dataset, memudahkan proses penyaringan data, serta memberikan konteks terhadap sumber komentar yang dianalisis (Pulungan *et al*., 2026). Selain itu, metadata membantu memastikan bahwa setiap komentar yang digunakan dalam penelitian memiliki keterkaitan dengan topik kecerdasan buatan (AI) dalam bidang pendidikan (Azzahra & Printo, 2025). Metadata yang digunakan dalam penelitian ini disajikan pada Tabel 3.2.

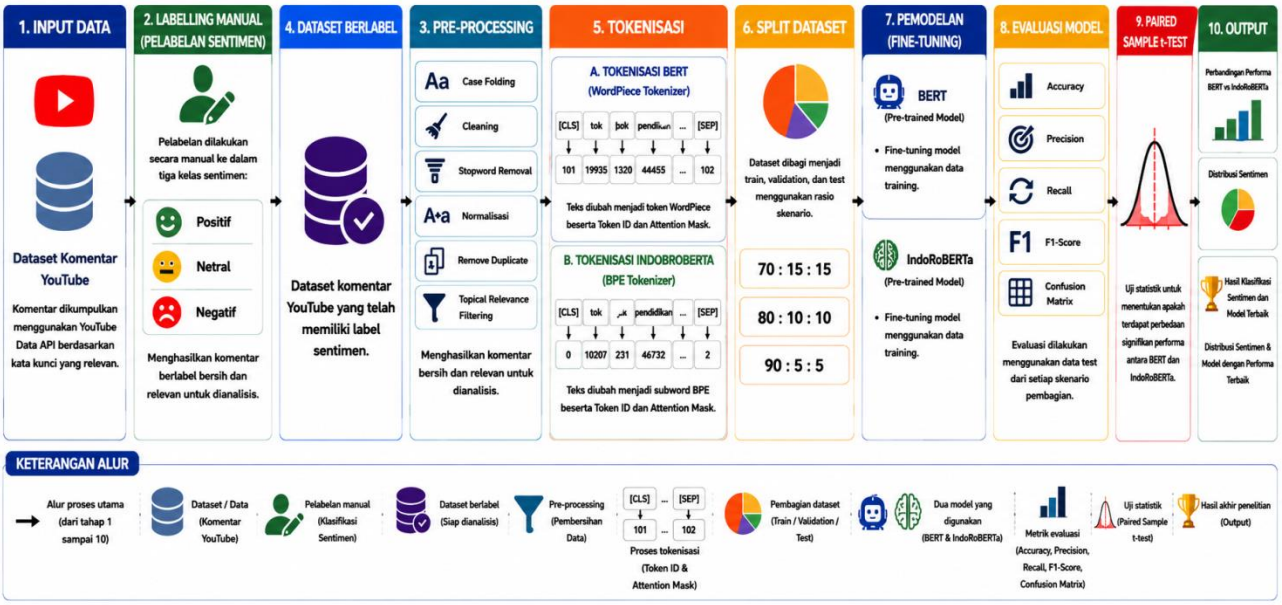
Tabel 3. 2 Metadata Dataset Komentar YouTube

Metadata	Keterangan
Comment	Isi komentar pengguna yang menjadi objek analisis sentimen
Published At	Waktu publikasi komentar
Video ID	Identitas unik video YouTube
Keyword	Kata kunci yang digunakan dalam proses pencarian data
Title	Judul video sumber komentar
Channel	Nama kanal YouTube yang mempublikasikan video

Berdasarkan Tabel 3.2, metadata digunakan untuk menjaga keterlacakan sumber data, mendukung proses *filtering* berdasarkan topik penelitian, serta memberikan informasi kontekstual terhadap komentar yang dianalisis. Informasi tersebut juga membantu proses identifikasi relevansi data sebelum dilakukan tahapan *pre-processing* dan klasifikasi sentimen.

3.2.2 Desain Sistem

Gambar 3.3 di bawah merupakan tahapan desain sistem yang akan digunakan dalam penelitian ini:



Gambar 3. 3 Desain Sistem

Berdasarkan Gambar 3.3, tahapan sistem terdiri atas sepuluh proses utama yaitu: (1) input data, (2) pelabelan manual, (3) *pre-processing*, (4) pembentukan dataset berlabel, (5) tokenisasi, (6) pembagian dataset, (7) pemodelan menggunakan BERT dan IndoRoBERTa, (8) evaluasi model, (9) uji statistik *paired sample t-test*, dan (10) output hasil klasifikasi sentimen.

a. Input Data

Data yang digunakan dalam penelitian ini berupa komentar pengguna YouTube yang membahas topik kecerdasan buatan (AI) dalam bidang pendidikan. Data diperoleh melalui proses *web scraping* menggunakan YouTube Data API v3 yang diakses melalui skrip *Python* dengan pendekatan pencarian berbasis kata kunci tematik, yaitu *AI pendidikan, kecerdasan buatan pendidikan, AI dalam pendidikan, ChatGPT pendidikan, AI di sekolah, dan etika AI dalam pendidikan.*

Berdasarkan kata kunci tersebut, sistem secara otomatis menelusuri video YouTube yang relevan dalam rentang waktu Agustus 2020 hingga Desember 2025 dan mengekstraksi komentar pengguna dari setiap video yang ditemukan.

Setiap komentar direpresentasikan sebagai satu unit data yang terdiri atas teks komentar serta metadata pendukung seperti waktu unggah komentar *(publishedAt)*, videoId, judul video, dan nama kanal. Seluruh data hasil ekstraksi disimpan dalam format XLS sebagai dataset awal yang akan diproses pada tahap berikutnya.

b. Pelabelan Manual Sentimen

Setelah proses pengumpulan dan penyaringan data selesai dilakukan, tahap berikutnya adalah pelabelan manual sentimen (*manual sentiment annotation*) terhadap setiap komentar yang telah dinyatakan relevan dengan topik penelitian. Pelabelan dilakukan oleh annotator dengan membaca dan menginterpretasikan isi komentar berdasarkan makna, konteks kalimat, serta kecenderungan opini yang disampaikan pengguna. Setiap komentar kemudian diklasifikasikan ke dalam tiga kategori sentimen, yaitu positif, netral, dan negatif.

Sentimen positif diberikan pada komentar yang menunjukkan dukungan, apresiasi, optimisme, atau pandangan yang baik terhadap pemanfaatan kecerdasan buatan (AI) dalam bidang pendidikan. Sentimen negatif diberikan pada komentar yang mengandung kritik, penolakan, kekhawatiran, atau persepsi negatif terhadap penggunaan AI. Sementara itu, sentimen netral diberikan pada komentar yang bersifat informatif, deskriptif, atau tidak menunjukkan kecenderungan opini yang jelas terhadap topik yang dibahas.

Proses pelabelan manual dilakukan karena kualitas label memiliki peran penting dalam pembentukan dataset klasifikasi sentimen yang valid dan representatif. Pada penelitian berbasis *supervised learning*, kualitas anotasi secara langsung memengaruhi performa model yang dilatih, sehingga pelabelan manual masih dianggap sebagai pendekatan yang mampu menghasilkan *ground truth* yang lebih akurat dibandingkan metode pelabelan otomatis (Atteveldt *et al*., 2021). Selain itu, penggunaan annotator manusia memungkinkan interpretasi konteks bahasa, ironi, serta variasi ekspresi yang sering ditemukan pada komentar media sosial dan sulit dikenali secara tepat oleh sistem otomatis (Rønningstad *et al*., 2024).

Hasil dari proses pelabelan ini berupa dataset berlabel yang selanjutnya digunakan sebagai data acuan (*ground truth*) dalam proses pelatihan, validasi, dan evaluasi model BERT maupun IndoRoBERTa. Dataset berlabel tersebut menjadi dasar bagi model untuk mempelajari pola linguistik dan karakteristik sentimen sehingga mampu melakukan klasifikasi sentimen secara lebih akurat pada data komentar YouTube berbahasa Indonesia. Proses anotasi manual juga banyak digunakan dalam penelitian analisis sentimen terkini karena terbukti mampu menghasilkan kualitas data yang lebih baik serta meningkatkan reliabilitas model klasifikasi berbasis Transformer (Lindén *et al*., 2023; Šmíd *et al*., 2024).

c. Dataset Berlabel

Dataset berlabel merupakan hasil keluaran dari proses pelabelan manual sentimen yang dilakukan terhadap komentar publik YouTube terkait kecerdasan buatan (AI) dalam bidang pendidikan. Setelah melalui tahap *pre-processing* dan dinyatakan relevan dengan topik penelitian, setiap komentar dianotasi secara manual oleh annotator ke dalam tiga kategori sentimen, yaitu positif, netral, dan negatif. Proses ini menghasilkan data berlabel yang berfungsi sebagai *ground truth* dalam penelitian, sehingga setiap komentar memiliki pasangan antara teks dan kelas sentimen yang sesuai. Keberadaan data berlabel sangat penting dalam pendekatan *supervised learning* karena digunakan sebagai dasar bagi model untuk mempelajari pola hubungan antara karakteristik teks dan kategori sentimen yang diwakilinya (Tiwari *et al*., 2023).

Dataset berlabel yang telah terbentuk selanjutnya digunakan pada tahapan *pre-processing*, tokenisasi, pembagian data (*dataset splitting*), pelatihan model (*training*), validasi, dan pengujian performa model. Kualitas pelabelan yang baik akan memengaruhi kemampuan model dalam mengenali pola linguistik dan konteks sentimen secara lebih akurat, terutama pada data media sosial yang mengandung bahasa informal, singkatan, serta variasi penulisan yang tinggi. Oleh karena itu, dataset berlabel pada penelitian ini menjadi landasan utama dalam proses *fine-tuning* model BERT dan IndoRoBERTa untuk menghasilkan klasifikasi sentimen yang akurat dan representatif terhadap persepsi publik

mengenai pemanfaatan AI dalam bidang pendidikan (D. Z. Abidin *et al*., 2025; Dhendra & Utomo, 2025).

d. *Pre-processing*

Tahap *pre-processing* bertujuan menyiapkan data teks agar dapat direpresentasikan secara optimal oleh model berbasis *transformer*, yaitu BERT dan IndoRoBERTa.

1. ***Case folding***

   Case folding mengubah seluruh teks menjadi huruf kecil untuk menyamakan representasi kata dan menghindari perbedaan akibat kapitalisasi, sehingga meningkatkan konsistensi data dan kinerja model klasifikasi sentimen (Siino *et al*., 2024).

2. ***Cleaning***

   Tahap *cleaning* dalam *preprocessing* teks menghapus elemen tidak relevan seperti URL, angka, simbol khusus, emotikon, dan tanda baca untuk mereduksi *noise* dan meningkatkan kualitas data sebelum analisis lanjutan. Proses ini penting untuk meningkatkan performa model NLP (Jim *et al*., 2024).

3. ***Stopword removal***

   Stopword removal adalah proses menghapus kata-kata umum tidak bermakna dari teks untuk mereduksi *noise* dan mengurangi dimensi fitur sehingga analisis NLP lebih efektif; teknik ini dapat dioptimalkan menggunakan *stop words list* seperti yang juga diekstraksi secara otomatis dari korpus besar Bahasa Indonesia untuk meningkatkan kualitas preprocessing teks (Achsan *et al*., 2023).

4. **Normalisasi teks**

   Normalisasi teks mengubah tulisan non-standar (singkatan, kata tidak baku, kesalahan ketik) ke bentuk baku agar konsisten secara linguistik sebelum analisis. Ini penting karena komentar YouTube bersifat informal; normalisasi meningkatkan kualitas representasi teks dan akurasi model NLP pada tugas klasifikasi dan analisis sentimen (Khan & Lee, 2021).

5. ***Remove duplicate comments***

Tahapan *remove duplicate comments* merupakan langkah penting dalam *pre-processing* untuk meningkatkan kualitas data dan akurasi model klasifikasi sentimen, karena komentar identik yang berulang dapat memicu *data leakage*, bias hasil, dan *overestimation* performa model — masalah yang perlu diatasi untuk keluaran lebih valid dan andal dalam NLP dan sentiment analysis (Mu et al., 2024).

6. ***Topical relevance filtering***

Penyaringan komentar digunakan untuk mempertahankan komentar yang benar-benar sesuai domain AI dalam pendidikan, dengan memeriksa kesesuaian komentar terhadap kata kunci tematik seperti *AI pendidikan*, *kecerdasan buatan pendidikan*, *ChatGPT pendidikan*, *AI di sekolah*, dan *etika AI dalam pendidikan*. Proses ini meningkatkan kualitas data dan akurasi klasifikasi sentimen berbasis BERT dan IndoRoBERTa (Santin & Rezende, 2025).

Tabel 3. 3 Contoh Tahapan *Pre-processing* Data Komentar YouTube

Tahapan	Contoh Sebelum	Contoh Sesudah
Case folding	"AI Sangat Menarik!"	"ai sangat menarik!"
Cleaning	"Cek link ini: https://youtube.com #AI123"	"cek link ini ai"
Stopword removal	"Saya sangat menyukai AI karena membantu belajar"	"menyukai ai membantu belajar"
Normalisasi teks	"Gue pgn ptn AI tp gk ngerti"	"saya ingin pelajari AI tapi tidak mengerti"
Remove duplicate comments	"AI keren!" "AI keren!"	"AI keren!"
Topical relevance filtering	"Film tadi malam seru"	*(dihapus karena tidak relevan dengan AI)*

Berdasarkan Tabel 3.3, tahapan *pre-processing* menunjukkan proses transformasi data komentar dari bentuk mentah menjadi lebih bersih dan terstruktur sebelum dianalisis. Setiap langkah memiliki fungsi spesifik, mulai dari

case folding untuk menyeragamkan huruf, *cleaning* untuk menghapus elemen tidak relevan, *stopword removal* untuk mengurangi kata umum yang tidak bermakna, hingga normalisasi teks untuk memperbaiki bahasa tidak baku. Selain itu, dilakukan penghapusan komentar duplikat guna menghindari bias data, serta *topical relevance filtering* untuk memastikan hanya komentar yang sesuai dengan topik kecerdasan buatan dalam pendidikan yang dipertahankan. Contoh sebelum dan sesudah pada tabel menunjukkan bahwa setiap tahapan secara bertahap meningkatkan kualitas teks sehingga lebih representatif dan siap digunakan dalam proses klasifikasi sentimen.

e. Tokenisasi

Tokenisasi merupakan tahap mengubah teks hasil *pre-processing* menjadi representasi numerik yang dapat diproses oleh model *transformer*. Pada penelitian ini, proses tokenisasi dilakukan menggunakan tokenizer yang sesuai dengan masing-masing model, yaitu *WordPiece Tokenizer* pada BERT dan *Byte Pair Encoding* (BPE) Tokenizer pada IndoRoBERTa. Melalui proses ini, setiap komentar dipecah menjadi unit *subword*, kemudian dikonversi menjadi *token ID* dan *attention mask* sebagai masukan (*input*) model. Penggunaan tokenisasi berbasis *subword* bertujuan untuk mengurangi permasalahan *out-of-vocabulary* (OOV), mempertahankan informasi linguistik, serta meningkatkan kemampuan model dalam memahami konteks bahasa secara lebih akurat. Penelitian terkini menunjukkan bahwa segmentasi *subword* memiliki peran penting dalam meningkatkan kualitas representasi bahasa dan performa model *transformer* pada berbagai tugas *Natural Language Processing* (NLP), termasuk klasifikasi sentimen (Samuel & Øvrelid, 2023).

1) Tokenisasi BERT

Pada model BERT, tokenisasi dilakukan menggunakan *WordPiece Tokenizer* yang memecah teks menjadi unit *subword* dan menambahkan token khusus [CLS] serta [SEP]. Hasil tokenisasi kemudian dikonversi menjadi *token ID* dan *attention mask* sebelum diproses oleh lapisan *Transformer Encoder* untuk menghasilkan representasi kontekstual yang digunakan dalam klasifikasi sentimen. Pendekatan *WordPiece* memungkinkan model mengenali kata yang

jarang muncul maupun variasi morfologi kata dengan tetap mempertahankan hubungan semantik antar token dalam kalimat (Geni *et al*., 2023).

2) Tokenisasi IndoRoBERTa

Pada model IndoRoBERTa, tokenisasi dilakukan menggunakan *Byte Pair Encoding* (BPE) *Tokenizer* yang membagi teks menjadi unit *subword* berdasarkan frekuensi kemunculan karakter dan pasangan kata dalam korpus pelatihan. Selanjutnya, hasil tokenisasi dikonversi menjadi *token ID* dan *attention mask* sebagai masukan ke dalam model IndoRoBERTa. Metode BPE dinilai lebih fleksibel dalam menangani variasi kosakata, bahasa informal, serta kata-kata baru yang banyak ditemukan pada komentar media sosial berbahasa Indonesia, sehingga dapat meningkatkan kualitas representasi kontekstual yang dihasilkan model (Samuel & Øvrelid, 2023; Hou *et al*., 2023).

f. Pembagian Dataset

Setelah tahap tokenisasi, dataset dibagi menjadi training set, validation set, dan test set menggunakan stratified split untuk menjaga proporsi kelas sentimen (positif, netral, negatif) tetap seimbang pada setiap subset. Teknik ini diterapkan untuk meminimalkan risiko ketidakseimbangan kelas yang dapat memengaruhi kinerja model (Tzimiris *et al*., 2025).

Penelitian ini menggunakan tiga skenario pembagian data untuk menguji pengaruh variasi proporsi data pelatihan terhadap performa model BERT dan IndoRoBERTa, yaitu 70% training, 15% validation, 15% test; 80% training, 10% validation, 10% test; serta 90% training, 5% validation, 5% test. Data training digunakan untuk *fine-tuning* model, data validation untuk pemantauan dan penentuan parameter optimal, sedangkan data test untuk evaluasi generalisasi model.

Skenario pembagian dataset tersebut disajikan pada Tabel 3.4 berikut.

Tabel 3. 4 Pembagian Dataset

Skenario	Training Set	Validation Set	Test Set
Skenario 1	70%	15%	15%
Skenario 2	80%	10%	10%
Skenario 3	90%	5%	5%

Berdasarkan Tabel 3.4, variasi pembagian dataset digunakan untuk memperoleh konfigurasi data yang menghasilkan performa klasifikasi terbaik pada model BERT dan IndoRoBERTa. Hasil evaluasi dari masing-masing skenario kemudian dibandingkan menggunakan metrik akurasi, presisi, *recall*, dan *F1-score* untuk menentukan model yang paling optimal dalam melakukan klasifikasi sentimen komentar YouTube terkait kecerdasan buatan dalam bidang pendidikan.

g. Model BERT dan IndoRoBERTa

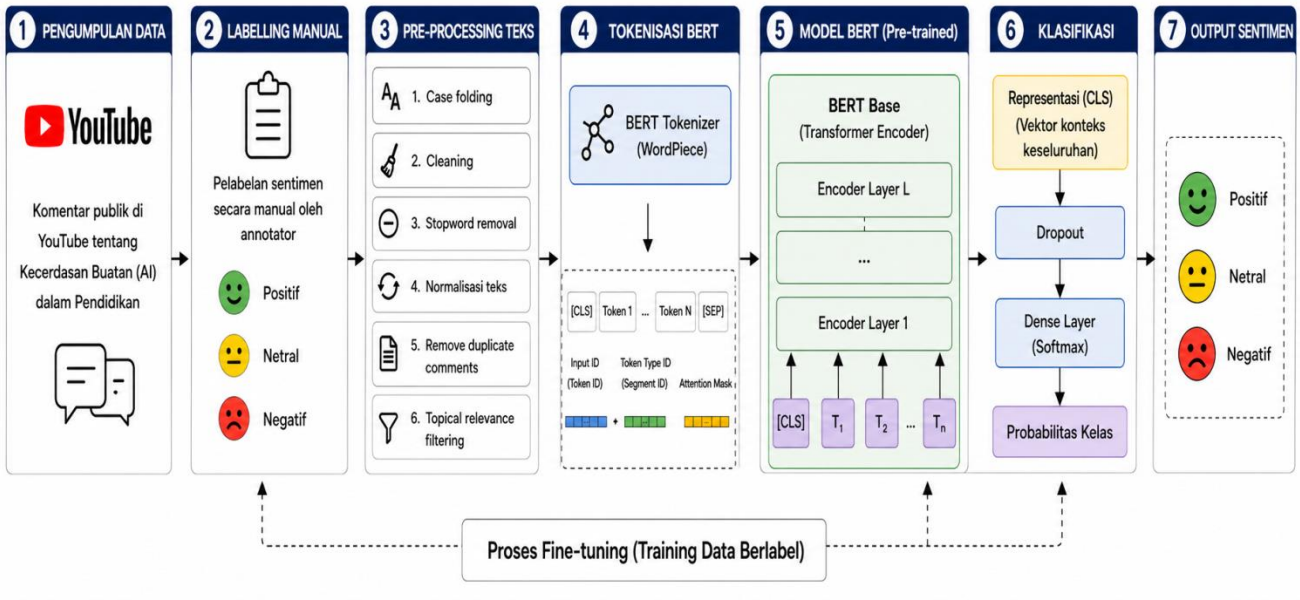
Klasifikasi sentimen dalam penelitian ini menggunakan dua model berbasis Transformer, yaitu BERT dan IndoRoBERTa. Kedua model memanfaatkan mekanisme self-attention untuk menghasilkan representasi kontekstual teks yang mendalam sehingga mampu memahami hubungan antar kata dalam suatu kalimat. Perbedaan utama keduanya terletak pada metode tokenisasi, korpus pelatihan, dan strategi pre-training yang digunakan.

1. BERT

BERT merupakan model berbasis *Transformer encoder* yang digunakan untuk membangun representasi teks secara kontekstual dua arah (*bidirectional*) (Gardazi *et al.*, 2025). Dalam penelitian ini, BERT digunakan untuk melakukan klasifikasi sentimen terhadap komentar publik YouTube terkait kecerdasan buatan dalam pendidikan.

a. Alur Proses Klasifikasi Bert

BERT digunakan untuk melakukan klasifikasi sentimen terhadap komentar publik di YouTube terkait kecerdasan buatan dalam pendidikan. Alur proses dimulai dari input data hingga menghasilkan output berupa label sentimen, sebagaimana ditunjukkan pada Gambar 3.4.



Gambar 3. 4 Alur Proses Klasifikasi BERT

Berdasarkan Gambar 3.4, tahapan pertama adalah pengumpulan data berupa komentar publik yang relevan dengan topik penelitian. Selanjutnya, data tersebut melalui proses pelabelan manual oleh annotator untuk mengklasifikasikan sentimen ke dalam kategori positif, netral, dan negatif. Setelah itu, dilakukan tahap *pre-processing* teks yang mencakup *case folding*, *cleaning*, *stopword removal*, normalisasi teks, penghapusan data duplikat, serta penyaringan relevansi topik guna meningkatkan kualitas data.

Tahap berikutnya adalah tokenisasi menggunakan BERT *Tokenizer* dengan metode *WordPiece*, yang mengubah teks menjadi token-token numerik serta menambahkan token khusus seperti [CLS] dan [SEP]. Token yang dihasilkan kemudian diproses dalam model BERT *pre-trained* yang terdiri dari beberapa lapisan encoder Transformer untuk menghasilkan representasi kontekstual.

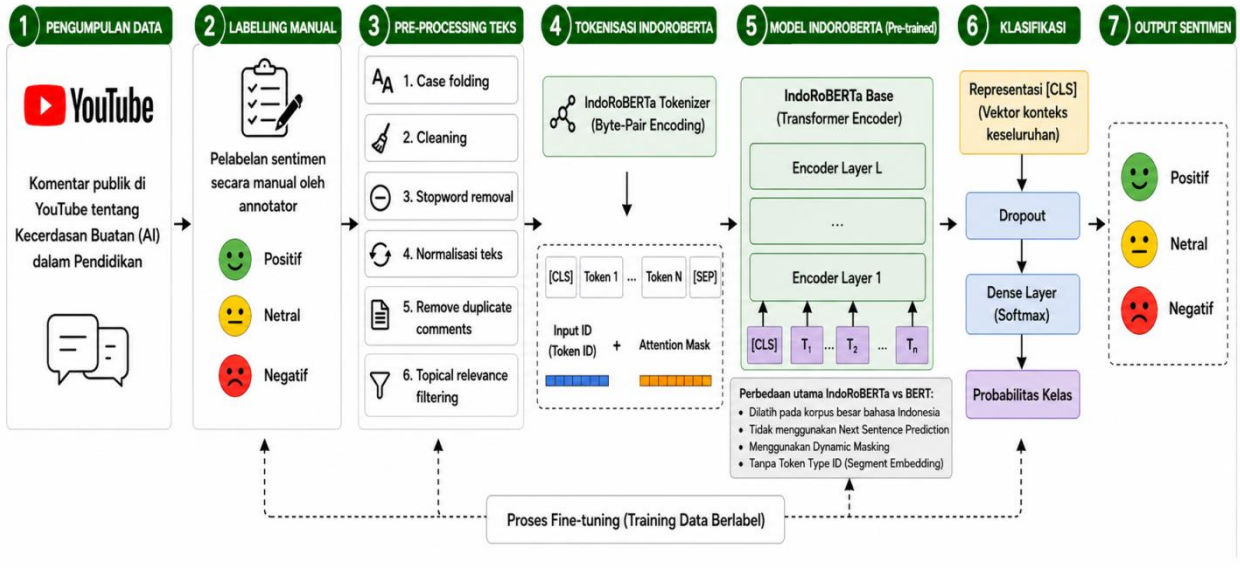
Selanjutnya, representasi dari token [CLS] digunakan dalam tahap klasifikasi melalui lapisan dropout dan dense layer dengan fungsi aktivasi *softmax* untuk menghasilkan probabilitas masing-masing kelas. Hasil akhir dari proses ini adalah output sentimen yang dikategorikan ke dalam kelas positif, netral, dan negatif.

2). IndoRoBERTa

IndoRoBERTa merupakan pengembangan dari model RoBERTa yang telah diadaptasi untuk bahasa Indonesia. Model ini dilatih menggunakan korpus besar berbahasa Indonesia sehingga lebih mampu menangkap karakteristik linguistik lokal dibandingkan BERT umum (Yulianti *et al.*, 2024; Nur *et al.*, 2025).

a. Alur Proses Klasifikasi IndoRoBERTa

IndoRoBERTa digunakan untuk klasifikasi sentimen dengan *pipeline* serupa, namun dengan perbedaan pada *pre-training* dan tokenisasi. Alur proses ditunjukkan pada Gambar 3.5.



Gambar 3. 5 Alur Proses Klasifikasi IndoRoBERTa

Berdasarkan Gambar 3.5, tahapan pertama dimulai dari pengumpulan data berupa komentar publik yang relevan dengan topik penelitian. Data yang telah dikumpulkan kemudian melalui proses pelabelan manual oleh annotator untuk mengelompokkan sentimen ke dalam kategori positif, netral, dan negatif. Tahap ini berfungsi sebagai dasar dalam penyediaan data berlabel untuk proses pelatihan model.

Selanjutnya, data diproses pada tahap *pre-processing* teks yang meliputi *case folding*, *cleaning*, *stopword removal*, normalisasi teks, penghapusan komentar duplikat, serta penyaringan relevansi topik. Tahapan ini bertujuan untuk

meningkatkan kualitas data sehingga lebih siap digunakan dalam proses pemodelan.

Tahap berikutnya adalah tokenisasi menggunakan IndoRoBERTa *Tokenizer* dengan pendekatan *Byte-Pair Encoding* (BPE). Pada tahap ini, teks dipecah menjadi *subword* yang kemudian dikonversi ke dalam bentuk token ID serta dilengkapi dengan *attention mask*. Token khusus seperti [CLS] digunakan sebagai representasi keseluruhan teks dalam proses klasifikasi.

Token yang telah diproses kemudian dimasukkan ke dalam model IndoRoBERTa *pre-trained* yang berbasis arsitektur *Transformer encoder*. Model ini terdiri dari beberapa lapisan *encoder* yang bertugas menghasilkan representasi kontekstual dari setiap token. Berbeda dengan BERT, IndoRoBERTa tidak menggunakan mekanisme *Next Sentence Prediction*, melainkan mengandalkan teknik *dynamic masking* serta pelatihan pada korpus besar berbahasa Indonesia untuk meningkatkan performa pemahaman bahasa.

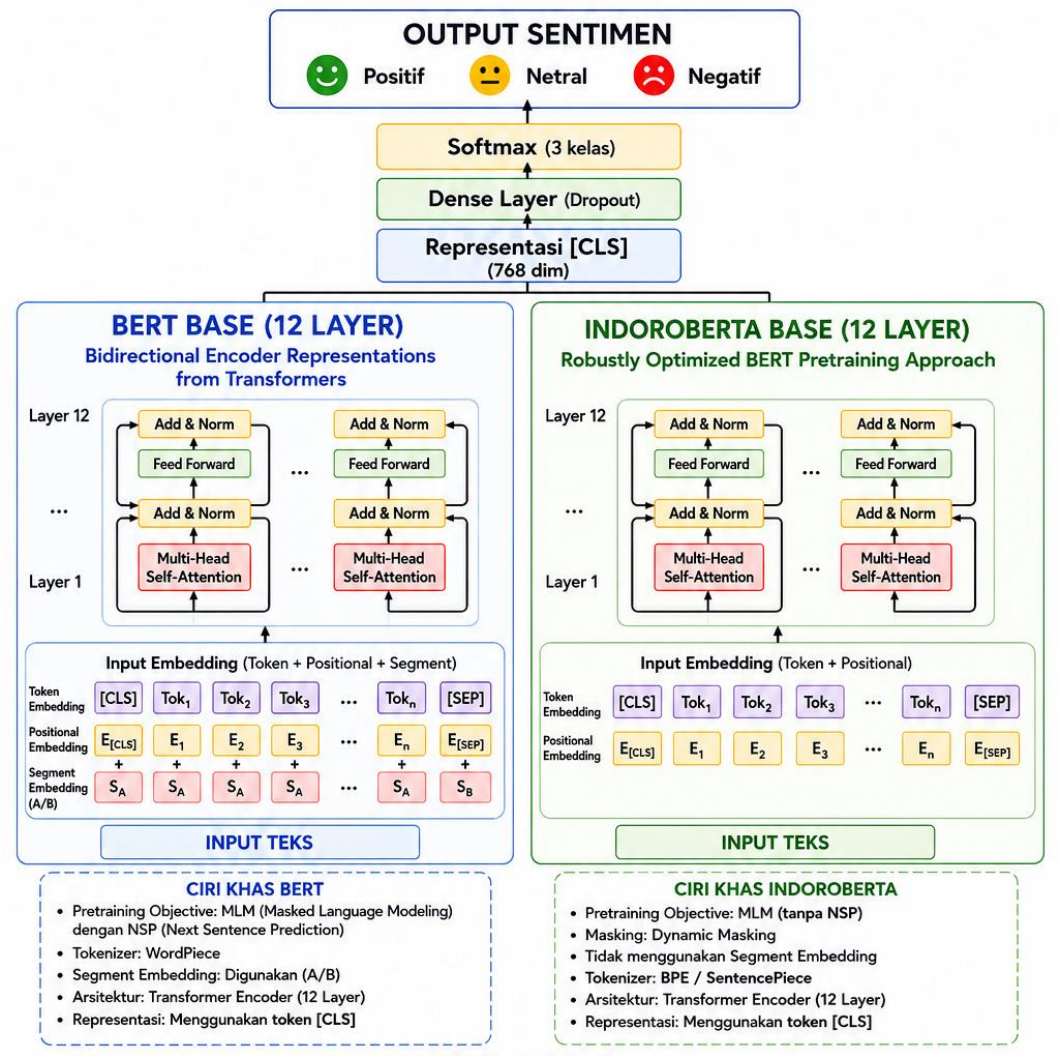
Pada tahap klasifikasi, representasi dari token [CLS] diteruskan ke lapisan *dropout* untuk mengurangi *overfitting*, kemudian ke *dense layer* dengan fungsi aktivasi *softmax* guna menghasilkan probabilitas masing-masing kelas sentimen. Hasil akhir dari proses ini berupa *output* sentimen yang dikategorikan ke dalam kelas positif, netral, dan negatif.

Dengan demikian, Gambar 3.6 menggambarkan secara sistematis alur kerja algoritma IndoRoBERTa mulai dari pengolahan data mentah hingga menghasilkan klasifikasi sentimen berbasis representasi kontekstual yang mendalam.

3. Perbandingan Arsitektur BERT dan IndoRoBERTa

BERT dan IndoRoBERTa merupakan model berbasis *Transformer Encoder* yang digunakan dalam penelitian ini untuk melakukan klasifikasi sentimen pada komentar publik *YouTube* terkait kecerdasan buatan dalam bidang pendidikan. Kedua model memanfaatkan mekanisme *self-attention* untuk menghasilkan representasi kontekstual teks secara dua arah sehingga mampu memahami hubungan antar kata dalam suatu kalimat. Meskipun memiliki struktur dasar yang serupa, terdapat beberapa perbedaan pada metode tokenisasi,

representasi *embedding*, serta strategi *pre-training* yang digunakan. Perbandingan arsitektur kedua model ditunjukkan pada Gambar 3.6.



Gambar 3. 6 Perbandingan Arsitektur BERT dan IndoRoBERTa

Berdasarkan Gambar 3.6, BERT dan IndoRoBERTa memiliki arsitektur dasar yang sama, yaitu *Transformer Encoder* yang terdiri atas 12 lapisan *encoder*. Kedua model memanfaatkan mekanisme *self-attention* untuk membangun representasi kontekstual teks dan menghasilkan representasi akhir yang digunakan dalam proses klasifikasi sentimen. Setelah melalui seluruh lapisan *encoder*, representasi token klasifikasi diproses melalui *dense layer* dan fungsi aktivasi *softmax* untuk menghasilkan probabilitas kelas sentimen positif, netral, dan negatif.

Perbedaan utama kedua model terletak pada tahap tokenisasi, representasi *embedding*, dan strategi *pre-training*. BERT menggunakan *WordPiece tokenizer*, *segment embedding*, serta *objective pre-training* berupa *Masked Language Modeling* (MLM) dan *Next Sentence Prediction* (NSP). Sebaliknya, IndoRoBERTa menggunakan *SentencePiece* berbasis *Byte Pair Encoding* (BPE), tidak menggunakan *segment embedding*, serta hanya menerapkan MLM dengan *dynamic masking* tanpa NSP. Perbedaan tersebut menjadikan IndoRoBERTa lebih optimal dalam memproses teks berbahasa Indonesia.

1) *Input Embedding*

Tahap pertama adalah pembentukan representasi numerik dari setiap token yang akan diproses oleh *Transformer encoder*.

BERT

Pada BERT, *embedding* diperoleh dari penjumlahan tiga komponen, yaitu token *embedding*, *positional embedding*, dan *segment embedding*.

$$E_i = E_i^{token} + E_i^{position} + E_i^{segment} \quad (3.1)$$

Keterangan:

- E_i : *embedding* token ke-*i*
- E_i^{token} : representasi token
- $E_i^{position}$: informasi posisi token
- $E_i^{segment}$: penanda segmen kalimat A atau B

Sesuai Gambar 3.6, BERT menggunakan *segment embedding* (A/B) untuk membedakan dua kalimat pada tahap *pre-training*.

IndoRoBERTa

IndoRoBERTa tidak menggunakan *segment embedding* sehingga *embedding* dibentuk hanya dari *token embedding* dan *positional embedding*.

$$E_i = E_i^{token} + E_i^{position} \quad (3.2)$$

Keterangan:

- E_i : *embedding* token ke-*i*
- E_i^{token} : representasi token
- $E_i^{position}$: informasi posisi token

Hal ini terlihat pada bagian *Input Embedding* (Token + Positional) pada arsitektur IndoRoBERTa.

2) *Transformer Encoder*

Setelah *embedding* terbentuk, token diproses oleh *Transformer Encoder* yang terdiri atas *Multi-Head Self-Attention*, *Feed Forward Network*, serta *Add & Norm*.

Query, Key, dan Value

Setiap token terlebih dahulu diproyeksikan menjadi *Query*, *Key*, dan *Value*.

$$Q = EW^Q, \quad K = EW^K, \quad V = EW^V \tag{3.3}$$

Keterangan:

- Q: *Query* (pencarian informasi)
- K: *Key* (representasi referensi)
- V: *Value* (informasi yang dibawa)
- W^Q, W^K, W^V : matriks bobot yang dipelajari

3) *Multi-Head Self-Attention*

Komponen utama *Transformer* adalah mekanisme *self-attention* yang menentukan hubungan antar token dalam kalimat.

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}_k}\right) \text{V} \tag{3.4}$$

Keterangan:

- QK^T: skor kesamaan antar token
- d_k : faktor normalisasi
- *softmax*: normalisasi bobot perhatian

Tahap ini memungkinkan model memahami hubungan kontekstual antar kata dalam kalimat.

4) *Multi-Head Attention*

Untuk menangkap berbagai jenis hubungan kata secara simultan digunakan beberapa *attention head*.

$$head_i = Attention\,(QW_i^Q, KW_i^K, VW_i^V) \tag{3.5}$$

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h)W^O \quad (3.6)$$

Keterangan:

- $head_i$: *attention* ke-i
- h : jumlah head attention
- W^O: bobot output

Pada Gambar 3.6 komponen ini ditunjukkan oleh blok *Multi-Head Self-Attention*.

5) *Feed Forward Network (FFN)*

Output *attention* kemudian diproses melalui jaringan *feed forward*.

$$FFN(x) = \max(0, xW_1 + b_1)\, W_2, + b_2 \quad (3.7)$$

Keterangan:

- x: input dari attention
- W_1, W_2: bobotb_1, b_2
- b_1, b_2 : bias

Tahap ini berfungsi memperkaya representasi fitur yang dihasilkan *encoder*.

6) *Add & Norm*

Untuk menjaga stabilitas pelatihan digunakan *residual connection* dan *layer normalization*.

$$Output = LayerNorm(x + Sublayer(x)) \quad (3.8)$$

Keterangan:

- x : *input* awal
- $Sublayer(x)$: output *attention* atau FFN
- $LayerNorm$: proses normalisasi

Pada Gambar 3.6 komponen ini ditunjukkan oleh blok *Add & Norm* pada setiap lapisan *encoder*.

7) Representasi Token Klasifikasi

Setelah melewati seluruh *encoder*, representasi token [CLS] digunakan sebagai representasi keseluruhan teks.

$$T_{[CLS]} \quad (3.9)$$

Keterangan:

- Representasi global dokumen
- Dimensi representasi umumnya 768

Representasi ini menjadi masukan pada tahap klasifikasi.

8) *Dense Layer* dan *Softmax*

Representasi token [CLS] diproses pada lapisan klasifikasi.

$$h = Dropout\left(T_{[CLS]}\right) \quad (3.10)$$

$$z = Wh + b \quad (3.11)$$

$$\hat{y} = softmax\,(z) \quad (3.12)$$

Keterangan Rumus:

- $Dropout$: mengurangi *overfitting*
- W, b: parameter model
- $\hat{y}$: probabilitas kelas

Sesuai Gambar 3.6, probabilitas yang dihasilkan digunakan untuk menentukan kelas sentimen.

9) Output Sentimen

Tahap akhir menghasilkan salah satu dari tiga kelas sentimen:

- Positif
- Netral
- Negatif

Kelas dengan nilai probabilitas tertinggi dipilih sebagai hasil klasifikasi akhir.

Berdasarkan Gambar 3.6, BERT dan IndoRoBERTa memiliki struktur *Transformer Encoder* yang sama sehingga mekanisme pembentukan representasi kontekstual dilakukan dengan cara yang serupa. Perbedaan utama terletak pada penggunaan *tokenizer*, *segment embedding*, *objective pre-training*, dan strategi *masking*. BERT menggunakan *WordPiece*, *segment embedding*, serta kombinasi *Masked Language Modeling* (MLM) dan *Next Sentence Prediction* (NSP), sedangkan IndoRoBERTa menggunakan *SentencePiece* berbasis *Byte Pair Encoding* (BPE), tidak menggunakan *segment embedding*, serta menerapkan MLM dengan *dynamic masking*. Ringkasan perbedaan kedua arsitektur disajikan pada Tabel 3.5.

Tabel 3. 5 Perbedaan Arsitektur BERT dan IndoRoBERTa

Komponen Arsitektur	BERT	IndoRoBERTa
Tokenizer	*WordPiece Tokenizer*	*SentencePiece Tokenizer* (BPE)
Input Embedding	*Token Embedding + Position Embedding + Segment Embedding*	*Token Embedding + Position Embedding*
Segment Embedding	Digunakan	Tidak digunakan
Pre-training Objective	MLM + NSP	MLM
Strategi *Masking*	*Static Masking*	*Dynamic Masking*
Data *Pre-training*	Korpus multibahasa	Korpus bahasa Indonesia
Representasi Input	Menggunakan informasi segmen kalimat	Tidak menggunakan informasi segmen kalimat
Fokus Optimasi	Pemahaman bahasa umum	Pemahaman bahasa Indonesia

3.2.3 Skenario Uji Coba

Skenario uji coba dalam penelitian ini dirancang untuk mengevaluasi performa model BERT dan IndoRoBERTa dalam mengklasifikasikan sentimen komentar publik *YouTube* terkait kecerdasan buatan (*Artificial Intelligence*/*AI*) dalam bidang pendidikan. Proses pengujian dilakukan melalui empat tahapan utama, yaitu pembagian dataset, pengujian *baseline*, proses *fine-tuning* model, serta eksperimen terhadap variasi pembagian data.

1. Pembagian Dataset (*Splitting Dataset*)

Dataset dibagi menjadi tiga subset, yaitu *training set*, *validation set*, dan *test set* menggunakan teknik *stratified split* untuk menjaga proporsi distribusi kelas sentimen (positif, netral, dan negatif) agar tetap seimbang pada setiap subset (Tzimiris *et al*., 2025).

Penelitian ini menggunakan tiga skenario pembagian data, yaitu:

- Skenario 1: 70% *training*, 15% *validation*, 15% *test*
- Skenario 2: 80% *training*, 10% *validation*, 10% *test*
- Skenario 3: 90% *training*, 5% *validation*, 5% *test*

Pada setiap skenario, *training set* digunakan untuk melatih model, *validation set* digunakan untuk memantau performa selama proses pelatihan, sedangkan *test set* digunakan untuk evaluasi akhir model terhadap data yang belum pernah dilihat sebelumnya (Joseph *et al*., 2022; Huo *et al*., 2023).

2. Pengujian Baseline

Sebelum dilakukan *fine-tuning*, model BERT dan IndoRoBERTa terlebih dahulu diuji dalam kondisi *baseline* menggunakan bobot *pre-trained* yang tersedia. Pada tahap ini, model digunakan untuk melakukan klasifikasi sentimen tanpa proses pelatihan ulang pada dataset penelitian. Pengujian *baseline* bertujuan untuk memperoleh gambaran performa awal model serta menjadi acuan dalam mengukur peningkatan kinerja setelah dilakukan *fine-tuning*.

Evaluasi dilakukan menggunakan *test set* dengan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil pengujian *baseline* kemudian dibandingkan dengan hasil setelah *fine-tuning* untuk mengetahui efektivitas proses pelatihan ulang terhadap kemampuan model dalam memahami konteks sentimen pada komentar publik *YouTube*.

3. Fine-Tuning Model

Setelah memperoleh hasil *baseline*, model BERT dan IndoRoBERTa dilatih ulang (*fine-tuned*) menggunakan *training set*. Pada tahap ini dilakukan penyesuaian *hyperparameter* seperti *learning rate*, *batch size*, dan jumlah *epoch* untuk memperoleh performa yang optimal sesuai karakteristik data penelitian.

Selama proses pelatihan, performa model dievaluasi pada setiap *epoch* menggunakan *validation set* melalui metrik *validation loss*, *accuracy*, *precision*, *recall*, dan *F1-score* (Luca *et al.*, 2021). Evaluasi ini bertujuan untuk memantau proses pembelajaran model sekaligus mendeteksi potensi *overfitting*.

Setelah proses *fine-tuning* selesai, model terbaik dievaluasi menggunakan *test set* untuk memperoleh performa akhir yang bersifat objektif karena data uji tidak digunakan selama proses pelatihan. Pendekatan ini umum digunakan pada model berbasis *Transformer* untuk memastikan hasil evaluasi yang valid dan dapat digeneralisasikan (Azzouzy *et al*., 2025).

4. Perbandingan Hasil Eksperimen

Tahap akhir dilakukan dengan membandingkan hasil pengujian *baseline* dan *fine-tuning* pada masing-masing model serta pada setiap skenario pembagian data. Perbandingan dilakukan berdasarkan nilai *accuracy*, *precision*, *recall*, dan *F1-score* untuk mengidentifikasi model dan konfigurasi data yang memberikan performa terbaik dalam klasifikasi sentimen komentar publik *YouTube* terkait kecerdasan buatan dalam bidang pendidikan.

3.2.4 Evaluasi

Evaluasi dilakukan baik pada tahap *baseline* maupun setelah proses *fine-tuning* untuk mengukur peningkatan performa model BERT dan IndoRoBERTa dalam mengklasifikasikan sentimen komentar publik di *YouTube*.

Evaluasi model dilakukan untuk menilai kinerja IndoRoBERTa dan BERT dalam mengklasifikasikan sentimen komentar YouTube ke dalam tiga kelas, yaitu positif, negatif, dan netral. Proses evaluasi menggunakan testing set yang terpisah dari data pelatihan dan validasi guna memastikan pengukuran kinerja yang objektif dan menghindari bias pembelajaran model. Kinerja model diukur menggunakan metrik evaluasi standar, yaitu accuracy, precision, recall, dan F1-score, yang telah banyak digunakan dalam penelitian analisis sentimen berbasis transformer untuk memberikan gambaran performa model yang komprehensif (Tzimiris et al., 2025).

1. Akurasi *(Accuracy)*

   Mengukur proporsi prediksi yang tepat dibandingkan dengan seluruh prediksi yang dihasilkan oleh model.

   Rumus:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.13}$$

   Keterangan:

   - TP *(True Positive)*: jumlah data suatu kelas sentimen yang diprediksi dengan benar

- TN *(True Negative)***:** jumlah data bukan kelas sentimen tersebut yang diprediksi dengan benar
- FP *(False Positive)*: jumlah data yang salah diprediksi sebagai kelas sentimen tertentu
- FN *(False Negative)*: jumlah data kelas sentimen tertentu yang gagal diprediksi

2. Presisi *(Precision)*

   Menilai tingkat ketepatan prediksi model terhadap suatu kelas sentimen tertentu, dengan membandingkan jumlah prediksi benar terhadap seluruh prediksi pada kelas tersebut.

   Rumus:

$$Precision = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Positive\ (FP)} \quad (3.14)$$

   Keterangan:

   - TP *(True Positive)*: prediksi benar pada kelas sentimen tertentu
   - FP *(False Positive)*: prediksi salah yang diklasifikasikan sebagai kelas tersebut

3. *Recall*

   Mengukur kemampuan model dalam mengenali seluruh data yang benar-benar termasuk dalam suatu kelas sentimen.

   Rumus:

$$Recall = \frac{TP}{TP + FN} \quad (3.15)$$

   Keterangan:

   - TP *(True Positive)*: data kelas sentimen tertentu yang berhasil diprediksi
   - FN *(False Negative)*: data kelas sentimen tertentu yang tidak terdeteksi oleh model

4. *F1-Score*

   Merupakan rata-rata harmonis antara *precision* dan *recall* yang digunakan

untuk menilai keseimbangan performa model, terutama ketika distribusi data antar kelas sentimen tidak seimbang.

Rumus:

$$F1 - Score = 2\,x\,\frac{Precision \times Recall}{Precision + Recall} \quad (3.16)$$

Keterangan:

- *Precision*: ketepatan prediksi model.
- *Recall:* kemampuan model mengenali data sebenarnya.
- *F1-Score*: keseimbangan antara precision dan recall.

Perhitungan metrik evaluasi dilakukan untuk masing-masing kelas sentimen (positif, negatif, dan netral). Selanjutnya, nilai kinerja keseluruhan dirangkum menggunakan *macro average* atau *weighted average* untuk memberikan gambaran performa model secara menyeluruh. Pendekatan evaluasi ini banyak digunakan dalam penelitian analisis sentimen berbasis transformer karena mampu menangkap nuansa konteks dan hubungan semantik antar kelas sentimen secara lebih (Azzouzy *et al*., 2025; Chi *et al*., 2025).

Selain itu, evaluasi juga diperkuat dengan *confusion matrix* untuk menganalisis distribusi prediksi benar dan salah pada setiap kelas secara lebih rinci (Fitroh & Hudaya, 2023; Rifaldi *et al.*, 2025). Hasil evaluasi ini menjadi dasar dalam membandingkan performa IndoRoBERTa dan BERT, serta dalam menentukan strategi *pre-processing* dan konfigurasi parameter yang paling optimal untuk klasifikasi sentimen komentar publik di YouTube.

3.2.5 Uji Statistik Paired Sample t-Test

Setelah dilakukan evaluasi kinerja model menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*, diperlukan pengujian statistik untuk mengetahui apakah perbedaan performa yang diperoleh antara model BERT dan IndoRoBERTa bersifat signifikan secara statistik. Pengujian statistik menjadi penting karena perbedaan nilai metrik evaluasi belum tentu menunjukkan adanya perbedaan kinerja yang nyata, melainkan dapat dipengaruhi oleh variasi data atau

proses pelatihan model. Oleh karena itu, penelitian ini menggunakan *paired sample t-test* sebagai metode untuk membandingkan performa kedua model secara kuantitatif dan objektif (Nariya *et al*., 2023).

Paired sample t-test merupakan metode statistik parametrik yang digunakan untuk membandingkan dua kelompok data yang saling berpasangan. Dalam konteks penelitian ini, pasangan data diperoleh dari hasil pengujian model BERT dan IndoRoBERTa pada dataset yang sama, sehingga setiap nilai evaluasi dari satu model memiliki pasangan nilai evaluasi yang sesuai pada model lainnya. Pengujian dilakukan terhadap nilai *accuracy* dan *F1-score* yang diperoleh dari masing-masing eksperimen untuk mengetahui apakah terdapat perbedaan performa yang signifikan antara kedua model (Nariya *et al*., 2023).

Rumus *paired sample t-test* ditunjukkan pada Persamaan (3.17).

$$t = \frac{\bar{d}}{\delta_d / \sqrt{n}} \tag{3.17}$$

Keterangan:

- t = nilai statistik uji *t*
- $\bar{d}$ = rata-rata selisih (*mean difference*) antara pasangan data
- δ_d = simpangan baku (*standard deviation*) dari selisih pasangan data
- $\sqrt{n}$ = jumlah pasangan data yang diuji

Nilai statistik *t* digunakan untuk mengukur besar perbedaan rata-rata antara dua kelompok data relatif terhadap variasi perbedaannya. Semakin besar nilai statistik *t*, semakin kuat indikasi bahwa perbedaan yang diamati tidak terjadi secara kebetulan. Selanjutnya, nilai tersebut digunakan untuk memperoleh *p-value* yang menjadi dasar dalam pengambilan keputusan statistik (Nahm, 2022).

Hipotesis penelitian yang digunakan adalah sebagai berikut:

- H_0 (Hipotesis Nol): Tidak terdapat perbedaan signifikan antara performa BERT dan IndoRoBERTa.
- H_1 (Hipotesis Alternatif): Terdapat perbedaan signifikan antara performa BERT dan IndoRoBERTa.

Kriteria pengambilan keputusan pada penelitian ini menggunakan tingkat signifikansi (α\alphaα) sebesar 0,05. Jika nilai *p-value* < 0,05 maka H_0 ditolak dan H_1 diterima, yang menunjukkan bahwa terdapat perbedaan performa yang signifikan antara model BERT dan IndoRoBERTa. Sebaliknya, jika nilai *p-value* ≥ 0,05 maka H_0 gagal ditolak, sehingga perbedaan performa yang diperoleh tidak dapat dinyatakan signifikan secara statistik (Nahm, 2022; Nariya *et al*., 2023).

Penggunaan *paired sample t-test* dalam penelitian ini bertujuan untuk memberikan validasi statistik terhadap hasil evaluasi model. Dengan demikian, kesimpulan mengenai model yang memiliki performa lebih baik tidak hanya didasarkan pada perbandingan nilai *accuracy* dan *F1-score*, tetapi juga didukung oleh bukti statistik yang menunjukkan tingkat signifikansi perbedaan performa antar model. Pendekatan ini banyak direkomendasikan dalam penelitian *machine learning* untuk memastikan bahwa peningkatan performa model benar-benar mencerminkan keunggulan metode yang digunakan dan bukan semata-mata akibat variasi sampel pengujian (Nariya *et al*., 2023).

BAB IV
HASIL DAN PEMBAHASAN

4.1 Hasil Pengumpulan Data

Dataset dalam penelitian ini diperoleh dari komentar publik pada video YouTube yang membahas kecerdasan buatan (AI) dalam konteks pendidikan. Pemilihan YouTube sebagai sumber data didasarkan pada sifatnya sebagai platform media sosial terbuka dengan tingkat interaksi tinggi yang mencerminkan opini publik secara spontan dan tidak terstruktur (Bal *et al*., 2024; Mutegi *et al*., 2025).

Pengumpulan data dilakukan menggunakan YouTube Data API dengan pendekatan *keyword-based crawling* untuk memastikan relevansi data terhadap topik penelitian. Kata kunci yang digunakan mencakup berbagai variasi istilah terkait AI dalam pendidikan, yang disusun untuk menangkap keragaman diskursus publik.

Hasil *crawling* awal diperoleh sebanyak 11.667 komentar dari 168 video dan 146 kanal YouTube, lengkap dengan metadata seperti waktu publikasi, *video ID*, judul video, dan nama kanal. Selanjutnya dilakukan proses validasi data yang meliputi penghapusan duplikasi, data kosong, serta penyaringan berdasarkan rentang waktu penelitian.

Setelah proses validasi, diperoleh 10.834 komentar valid dari 167 video dan 145 kanal YouTube. Rentang waktu pengumpulan data ditetapkan mulai 31 Maret 2022 hingga 31 Desember 2025 untuk menangkap dinamika perkembangan opini publik terhadap kecerdasan buatan dalam pendidikan secara temporal.

Untuk memberikan gambaran umum mengenai karakteristik dataset yang digunakan dalam penelitian ini, ringkasan data disajikan pada Tabel 4.1 berikut.

Tabel 4. 1 Ringkasan Dataset

Deskripsi	Nilai
Total komentar terkumpul	11.667
Komentar valid	10.834
Total video	167
Total kanal	145

Berdasarkan Tabel 4.1, dapat diketahui bahwa dataset akhir yang digunakan dalam penelitian ini terdiri atas 10.834 komentar valid yang berasal dari 167 video dan 145 kanal YouTube. Jumlah data yang relatif besar ini memberikan keuntungan dalam konteks pembelajaran mesin *(machine learning)*, khususnya pada model *deep learning* berbasis *transformer*, karena semakin besar dan beragam data yang digunakan, maka semakin baik kemampuan model dalam mempelajari pola-pola linguistik yang kompleks dalam teks.

Selain itu, banyaknya jumlah video dan kanal yang terlibat menunjukkan bahwa data yang digunakan tidak hanya berasal dari satu sumber atau satu jenis konten saja, melainkan mencakup berbagai perspektif dan gaya penyampaian yang berbeda. Hal ini memperkuat validitas dataset sebagai representasi dari opini publik yang beragam terhadap isu kecerdasan buatan dalam pendidikan.

Rentang waktu data yang mencakup periode tahun 2022 hingga 2025 juga menunjukkan bahwa dataset yang digunakan bersifat aktual dan relevan dengan perkembangan terkini teknologi kecerdasan buatan. Periode tersebut merupakan masa di mana perkembangan AI, khususnya model generatif dan *chatbot* seperti ChatGPT, mengalami peningkatan signifikan dalam adopsi dan diskusi publik di berbagai sektor, termasuk dunia pendidikan. Oleh karena itu, dataset ini dianggap mampu merepresentasikan kondisi empiris terkini dari persepsi masyarakat terhadap AI.

Untuk memberikan gambaran yang lebih konkret mengenai bentuk data yang digunakan dalam penelitian ini, Tabel 4.2 berikut menyajikan contoh data komentar YouTube yang diperoleh dari proses *crawling* sebelum dilakukan tahap *pre-processing*.

Tabel 4. 2 Contoh Dataset Komentar YouTube (Data Mentah)

Comment	Published At	Video ID	keyword	Title	Channel
Hari ini AI masih bayi, kita tunggu si AI beranjak remaja 2030 akan sedahsyat apa dia...	2024-05-27	_FcHPrRY5zg	tantangan AI pendidikan	AI Sama Sekali Enggak Seperti yang Kalian Kira (ft. Maudy Ayunda)	Kok Bisa?
Sadarkah anda? AI dibuat oleh penguasa dunia agar manusia semakin bergantung...	2024-05-25	_FcHPrRY5zg	tantangan AI pendidikan	AI Sama Sekali Enggak Seperti yang Kalian Kira (ft. Maudy Ayunda)	Kok Bisa?
Indonesia butuh literasi digital dan AI sejak TK hingga mahasiswa	2024-05-21	_FcHPrRY5zg	tantangan AI pendidikan	AI Sama Sekali Enggak Seperti yang Kalian Kira (ft. Maudy Ayunda)	Kok Bisa?
Secanggih apapun robot, belum bisa secanggih manusia yang bisa punya anak	2024-05-18	_FcHPrRY5zg	tantangan AI pendidikan	AI Sama Sekali Enggak Seperti yang Kalian Kira (ft. Maudy Ayunda)	Kok Bisa?
…	…	…	…	…	…

Berdasarkan Tabel 4.2, dapat diketahui bahwa data yang ditampilkan merupakan sebagian kecil dari dataset komentar YouTube yang digunakan dalam penelitian ini. Data tersebut masih berada dalam bentuk mentah *(raw data)* yang diperoleh langsung dari proses *crawling* menggunakan YouTube Data API sebelum melalui tahap *pre-processing*.

Setiap baris data terdiri dari enam atribut utama, yaitu *Comment*, *Published At*, *Video ID*, *Keyword*, *Title*, dan *Channel*. Atribut *Comment* merepresentasikan teks opini pengguna YouTube yang menjadi objek utama dalam klasifikasi sentimen. Atribut *Published At* menunjukkan waktu publikasi komentar yang memungkinkan adanya analisis temporal terhadap pola opini publik. Sementara itu, *Video ID* berfungsi sebagai identifikasi unik video sumber komentar yang digunakan dalam proses pengambilan data.

Atribut *Keyword* menunjukkan kata kunci yang digunakan dalam proses pencarian video pada tahap *crawling*, sehingga dapat memberikan gambaran

konteks tematik dari data yang dikumpulkan. Selanjutnya, *Title* merepresentasikan judul video yang menjadi sumber komentar, sedangkan *Channel* menunjukkan nama kanal YouTube yang mempublikasikan video tersebut.

Berdasarkan isi komentar yang ditampilkan, dapat diamati bahwa data memiliki karakteristik bahasa yang bersifat informal dan tidak terstruktur. Hal ini ditunjukkan oleh penggunaan gaya bahasa bebas, kalimat opini yang bersifat subjektif, serta adanya perbedaan cara penyampaian antar pengguna. Sebagian komentar mencerminkan pandangan kritis terhadap perkembangan kecerdasan buatan, sebagian lainnya menunjukkan sikap optimis, sementara sebagian lain bersifat reflektif terhadap dampak sosial teknologi AI.

Keragaman bentuk dan isi komentar tersebut menunjukkan bahwa data memiliki variasi linguistik yang tinggi, sehingga memerlukan tahap *pre-processing* yang tepat sebelum digunakan dalam proses klasifikasi sentimen. Selain itu, keberadaan atribut tambahan seperti *Published At*, *Video ID*, *Keyword*, *Title*, dan *Channel* memberikan konteks tambahan yang penting untuk memahami sumber, waktu, serta konteks kemunculan opini publik.

Dengan demikian, tabel ini tidak hanya berfungsi sebagai representasi data awal, tetapi juga menggambarkan kompleksitas dan kekayaan informasi yang menjadi dasar dalam proses analisis sentimen menggunakan model berbasis transformer dalam penelitian ini.

4.2 Pelabelan Manual dan Statistik Dataset

Tahap pelabelan sentimen dilakukan terhadap seluruh komentar yang telah dikumpulkan untuk mengonversi data teks tidak terstruktur menjadi data terstruktur yang memiliki label sebagai *ground truth* dalam proses pembelajaran model klasifikasi. Dalam penelitian ini, setiap komentar dikategorikan ke dalam tiga kelas sentimen, yaitu negatif, netral, dan positif, berdasarkan polaritas makna yang terkandung di dalamnya (Atteveldt *et al*., 2021; Alaei *et al*., 2023).

Proses pelabelan dilakukan secara manual oleh dua validator yang bekerja secara independen guna meminimalkan bias subjektif serta meningkatkan reliabilitas hasil pelabelan. Penilaian dilakukan berdasarkan pemahaman terhadap

konteks kalimat, struktur bahasa, serta makna implisit dalam komentar (Frenda *et al*., 2025). Pendekatan ini penting mengingat karakteristik data media sosial yang cenderung tidak baku dan mengandung ambiguitas.

Kriteria pelabelan ditetapkan sebagai berikut: komentar yang menunjukkan dukungan atau pandangan optimis terhadap kecerdasan buatan dalam pendidikan dikategorikan sebagai sentimen positif, komentar yang mengandung kritik atau kekhawatiran dikategorikan sebagai sentimen negatif, sedangkan komentar yang bersifat deskriptif atau tidak menunjukkan opini yang jelas diklasifikasikan sebagai sentimen netral (Trisna & Jie, 2022; Ni & Ni, 2024).

Untuk menjaga konsistensi antar validator, digunakan metode *inter-annotator agreement* untuk mengevaluasi tingkat kesepakatan pelabelan (Martínez-miwa & Castelán, 2023; Lindén *et al*., 2023). Apabila terdapat perbedaan label, dilakukan diskusi hingga diperoleh kesepakatan akhir (Stefanovitch, 2023).

Untuk memberikan gambaran konkret mengenai penerapan kriteria pelabelan, sebagian contoh hasil pelabelan manual disajikan pada Tabel 4.3. Tabel ini menunjukkan bagaimana komentar diklasifikasikan ke dalam kategori sentimen berdasarkan konteks dan makna yang terkandung di dalamnya.

Tabel 4. 3 Contoh Hasil Pelabelan Manual

Comment	Published At	Video ID	keyword	Title	Channel	Label
Tidak ada yang bisa kalah kita sebagai manusia, karena manusia diciptakan Allah SWT sebagai makhluk sempurna	2025-12-24 12:13:18	Glvgjr-gPUQ	AI pendidikan	[FULL] KICK ANDY - PROF STELLA: OTAK VS AI	Kok Bisa?	1
Otak lebih pintar dari teknologi	2025-12-24 12:10:24	Glvgjr-gPUQ	AI pendidikan	[FULL] KICK ANDY - PROF	Kok Bisa?	2

Ya tetap otak lah, tidak mungkin ada AI tanpa ada otak yang beride	2025-11-21 23:35:22	Glvgjr-gPUQ	AI pendidikan	STELLA: OTAK VS AI [FULL] KICK ANDY - PROF STELLA: OTAK VS AI [FULL] KICK ANDY - PROF STELLA: OTAK VS AI	Kok Bisa?	1
Saya kira biasa saja, ternyata mampu menjelaskan apa sebenarnya AI	2025-09-10 05:27:47	Glvgjr-gPUQ	AI pendidikan	STELLA: OTAK VS AI [FULL] KICK ANDY - PROF STELLA: OTAK VS AI	Kok Bisa?	2
…	…	…	…	…	…	

Keterangan:
Label 0 = Negatif, Label 1 = Netral, Label 2 = Positif

Tabel 4.3 menyajikan contoh hasil pelabelan manual terhadap data komentar YouTube yang digunakan dalam penelitian ini, yang dilengkapi dengan atribut metadata seperti waktu publikasi (*published at*), identitas video (*video ID*), kata kunci (*keyword*), judul video (*title*), dan kanal (*channel*) untuk memberikan konteks sumber data. Kolom *comment* berisi teks yang menjadi objek analisis, sedangkan kolom *label* menunjukkan hasil klasifikasi sentimen oleh validator, yaitu 0 (negatif), 1 (netral), dan 2 (positif), yang ditentukan berdasarkan analisis makna, konteks, dan kecenderungan opini dalam komentar. Penyajian tabel ini bertujuan untuk memberikan ilustrasi empiris mengenai penerapan kriteria pelabelan, sehingga memperjelas keterkaitan antara karakteristik teks dan kategori sentimen sebelum dianalisis lebih lanjut pada distribusi dataset secara keseluruhan.

Berdasarkan keseluruhan proses pelabelan terhadap seluruh dataset, diperoleh distribusi jumlah data pada masing-masing kategori sentimen yang dirangkum dalam Tabel 4.4.

Tabel 4. 4 Distribusi Hasil Pelabelan Sentimen

Label	Kategori	Jumlah Data
0	Negatif	384
1	Netral	9.006
2	Positif	1.444
	Total	**10.834**

Berdasarkan hasil tersebut, terlihat bahwa distribusi kelas tidak seimbang, di mana kategori netral mendominasi dataset dengan jumlah yang jauh lebih besar dibandingkan kategori lainnya. Kondisi ini menunjukkan bahwa sebagian besar komentar bersifat informatif atau deskriptif tanpa kecenderungan opini yang kuat.

Sementara itu, kategori positif dan negatif memiliki jumlah yang relatif lebih kecil sehingga tergolong sebagai kelas minoritas. Ketidakseimbangan ini merupakan karakteristik umum dalam data media sosial, di mana distribusi opini tidak merata dan cenderung didominasi oleh konten non-opini (Jamil *et al*., 2024; Bal *et al*., 2024; Widyasari & Muljono, 2026).

Ketidakseimbangan kelas berpotensi menyebabkan model klasifikasi bias terhadap kelas mayoritas. Oleh karena itu, penelitian ini menerapkan metode *class weight* dalam proses pelatihan model, yaitu dengan memberikan bobot yang lebih besar pada kelas minoritas dan bobot yang lebih kecil pada kelas mayoritas (Fachrie *et al*., 2025; Carvalho *et al*., 2025).

Pendekatan ini dipilih karena mampu menjaga distribusi asli data tanpa melakukan modifikasi jumlah sampel, sehingga tetap merepresentasikan kondisi nyata dari data komentar publik sekaligus meningkatkan kemampuan model dalam mengenali kelas minoritas (Ritonga *et al*., 2025).

4.4 Hasil *Pre-processing*

Tahap *pre-processing* dilakukan untuk menyiapkan data teks sebelum klasifikasi sentimen. Proses ini bertujuan meningkatkan kualitas data dengan mengurangi *noise*, menyeragamkan format, serta mengekstraksi informasi yang relevan. Tahapan yang diterapkan meliputi *case folding*, *cleaning*, *stopword*

removal, normalisasi teks, penghapusan data duplikat, dan penyaringan relevansi topik.

4.4.1 *Case Folding*

Case folding merupakan tahap awal dalam *pre-processing* yang bertujuan menyeragamkan seluruh teks menjadi bentuk huruf kecil *(lowercase)* untuk mengurangi variasi penulisan yang tidak konsisten dalam data. Hasil dari proses *case folding* pada data komentar YouTube dalam penelitian ini ditunjukkan pada Tabel 4.5.

Tabel 4. 5 Hasil *Case Folding*

Comment	***Case Folding***
intinya ..rumus adalah rumus dan membantu manusia hanyalah data..saat informasi atau data kurang lengkap dia akan terus mencari...	intinya ..rumus adalah rumus dan membantu manusia hanyalah data..saat informasi atau data kurang lengkap dia akan terus mencari...
Tinggalin jejak disini Semoga sales masih aman walau ada AI COMMENT 2025	tinggalin jejak disini semoga sales masih aman walau ada ai comment 2025
AI itu penemuan terbesar manusia saya sih setuju aja koding nya di tambah...	ai itu penemuan terbesar manusia saya sih setuju aja koding nya di tambah...
CARA KERJA AI DONG KAK, SEHINGGA BISA MENGGAMBAR DAN MENJAWAB SEMUA PERTANYAAN KITA	cara kerja ai dong kak, sehingga bisa menggambar dan menjawab semua pertanyaan kita
Seberapa canggihpun A.I mereka tidak akan pernah bisa menembus ruang waktu.	seberapa canggihpun ai mereka tidak akan pernah bisa menembus ruang waktu

Berdasarkan Tabel 4.5, dapat dilihat bahwa proses *case folding* berhasil mengubah seluruh teks komentar menjadi huruf kecil secara konsisten. Perubahan ini terlihat jelas pada komentar yang sebelumnya menggunakan huruf kapital, seperti *"CARA KERJA AI DONG KAK"* yang diubah menjadi *"cara kerja ai dong kak"*.

Selain itu, istilah seperti *"AI"* juga dikonversi menjadi *"ai"* sehingga tidak terjadi perbedaan representasi dalam proses analisis selanjutnya. Dengan demikian, tahap *case folding* mampu menyederhanakan variasi penulisan tanpa mengubah makna teks, serta meningkatkan konsistensi data untuk tahap *pre-processing* berikutnya.

4.4.2 *Cleaning*

Setelah tahap *case folding*, data kemudian diproses pada tahap *cleaning* untuk menghilangkan karakter yang tidak relevan seperti tanda baca dan simbol yang dapat menimbulkan *noise* dalam analisis teks. Hasil dari proses *cleaning* ditunjukkan pada Tabel 4.6.

Tabel 4. 6 Hasil *Cleaning*

Case Folding	***Cleaning***
intinya ..rumus adalah rumus dan membantu manusia hanyalah data..saat informasi atau data kurang lengkap dia akan terus mencari...	intinya rumus adalah rumus dan membantu manusia hanyalah datasaat informasi atau data kurang lengkap dia akan terus mencari ai bisa berbahaya karna ai haus informasi membantu manusia adalah program atau data jdi prioritas ai hnya pada alogaritma ya rumusnya jdi ai sangat sangat bahaya
tinggalin jejak disini semoga sales masih aman walau ada ai comment 2025	tinggalin jejak disini semoga sales masih aman walau ada ai comment 2025
ai itu penemuan terbesar manusia saya sih setuju aja koding nya di tambah...	ai itu penemuan terbesar manusia saya sih setuju aja koding nya di tambah yang menjadi inteligensia menjadi super inteligensia luar biasa saya sih penasaran sehebat apa ai melampaui manusia wkwk
cara kerja ai dong kak, sehingga bisa menggambar dan menjawab semua pertanyaan kita	cara kerja ai dong kak sehingga bisa menggambar dan menjawab semua pertanyaan kita
seberapa canggihpun ai mereka tidak akan pernah bisa menembus ruang waktu	seberapa canggihpun ai mereka tidak akan pernah bisa menembus ruang waktu

Berdasarkan Tabel 4.6, dapat dilihat bahwa proses *cleaning* berhasil menghilangkan berbagai tanda baca dan karakter yang tidak diperlukan dalam teks. Sebagai contoh, penggunaan tanda titik berulang (“..”) pada kalimat dihilangkan sehingga menghasilkan teks yang lebih bersih dan mudah diproses. Selain itu, tanda koma pada kalimat seperti “cara kerja ai dong kak, sehingga bisa menggambar...” juga dihapus tanpa mengubah makna dasar dari kalimat tersebut.

Namun demikian, pada beberapa data terlihat bahwa proses *cleaning* juga menyebabkan penggabungan kata, seperti pada kata “datasaat” yang berasal dari “data..saat”. Hal ini menunjukkan bahwa meskipun *cleaning* efektif dalam menghilangkan *noise*, diperlukan perhatian lebih pada tahap selanjutnya, seperti normalisasi, untuk memastikan struktur kata tetap optimal.

Secara keseluruhan, tahap *cleaning* berperan penting dalam meningkatkan kualitas data dengan menghilangkan elemen yang tidak relevan, sehingga data menjadi lebih siap untuk tahap *pre-processing* berikutnya, yaitu *stopword removal* dan normalisasi teks.

4.4.3 *Stopword Removal*

Setelah tahap *cleaning*, data selanjutnya diproses pada tahap *stopword removal* untuk menghilangkan kata-kata umum yang tidak memiliki kontribusi signifikan terhadap makna sentimen sehingga fitur teks menjadi lebih informatif. Hasil dari proses *stopword removal* ditunjukkan pada Tabel 4.7.

Tabel 4. 7 Hasil *Stopword Removal*

Cleaning	***Stopword Removal***
intinya ..rumus adalah rumus dan membantu manusia hanyalah data..saat informasi atau data kurang lengkap dia akan terus mencari...	intinya rumus adalah rumus dan membantu manusia hanyalah datasaat informasi atau data kurang lengkap dia akan terus mencari ai bisa berbahaya karna ai haus informasi membantu manusia adalah program atau data jdi prioritas ai hnya pada alogaritma ya rumusnya jdi ai sangat sangat bahaya
tinggalin jejak disini semoga sales masih aman walau ada ai comment 2025	tinggalin jejak disini semoga sales masih aman walau ada ai comment 2025
ai itu penemuan terbesar manusia saya sih setuju aja koding nya di tambah...	ai itu penemuan terbesar manusia saya sih setuju aja koding nya di tambah yang menjadi inteligensia menjadi super inteligensia luar biasa saya sih penasaran sehebat apa ai melampaui manusia wkwk
cara kerja ai dong kak, sehingga bisa menggambar dan menjawab semua pertanyaan kita	cara kerja ai dong kak sehingga bisa menggambar dan menjawab semua pertanyaan kita
seberapa canggihpun ai mereka tidak akan pernah bisa menembus ruang waktu	seberapa canggihpun ai mereka tidak akan pernah bisa menembus ruang waktu

Berdasarkan Tabel 4.7, dapat dilihat bahwa proses *stopword removal* berhasil menghilangkan kata-kata umum yang tidak memiliki makna signifikan, seperti “dan”, “yang”, “akan”, dan “bisa”. Penghapusan kata-kata tersebut membuat teks menjadi lebih ringkas dan fokus pada kata-kata penting yang merepresentasikan makna utama.

Sebagai contoh, kalimat *“cara kerja ai dong kak sehingga bisa menggambar dan menjawab semua pertanyaan kita”* disederhanakan menjadi

"cara kerja ai dong kak menggambar menjawab semua pertanyaan". Proses ini menunjukkan bahwa informasi utama tetap dipertahankan meskipun jumlah kata berkurang.

Dengan demikian, tahap *stopword removal* berperan dalam mengurangi kompleksitas data serta meningkatkan kualitas fitur yang akan digunakan dalam proses klasifikasi sentimen.

4.4.4 Normalisasi Teks

Setelah tahap *stopword removal*, data selanjutnya diproses pada tahap normalisasi untuk menyetarakan kata tidak baku atau variasi penulisan ke dalam bentuk standar agar meningkatkan konsistensi linguistik data. Hasil dari proses normalisasi teks ditunjukkan pada Tabel 4.8.

Tabel 4. 8 Hasil Normalisasi Teks

Stopword Removed	***Normalized***
intinya rumus rumus membantu manusia hanyalah datasaat informasi data kurang lengkap terus mencari ai berbahaya karna ai haus informasi membantu manusia program data jdi prioritas ai hnya alogaritma rumusnya jdi ai sangat sangat bahaya	intinya rumus rumus membantu manusia hanyalah datasaat informasi data kurang lengkap terus mencari ai berbahaya karna ai haus informasi membantu manusia program data jdi prioritas ai hnya alogaritma rumusnya jdi ai sangat sangat bahaya
tinggalin jejak disini semoga sales aman ai comment 2025	tinggalin jejak disini semoga sales aman ai comment 2025
ai penemuan terbesar manusia sih setuju aja koding nya tambah menjadi inteligensia menjadi super inteligensia luar biasa sih penasaran sehebat apa ai melampaui manusia wkwk	ai penemuan terbesar manusia sih setuju aja koding nya tambah menjadi inteligensia menjadi super inteligensia luar biasa sih penasaran sehebat apa ai melampaui manusia tertawa
cara kerja ai dong kak menggambar menjawab semua pertanyaan	cara kerja ai dong kak menggambar menjawab semua pertanyaan
seberapa canggihpun ai pernah menembus ruang waktu	seberapa canggihpun ai pernah menembus ruang waktu

Berdasarkan Tabel 4.8, proses normalisasi teks menunjukkan adanya penyederhanaan kata tidak baku menjadi bentuk yang lebih standar. Salah satu contoh terlihat pada kata *"wkwk"* yang diubah menjadi *"tertawa"*, sehingga makna emosional dalam teks tetap terjaga namun dalam bentuk yang lebih formal.

Namun, pada beberapa data tidak terjadi perubahan signifikan karena kata-kata yang digunakan sudah dalam bentuk yang cukup baku. Hal ini menunjukkan

bahwa proses normalisasi bersifat selektif dan hanya diterapkan pada kata-kata yang memerlukan penyesuaian.

Secara keseluruhan, tahap normalisasi teks berperan dalam meningkatkan konsistensi linguistik data serta membantu model dalam memahami makna teks dengan lebih baik, khususnya pada data yang berasal dari media sosial yang cenderung tidak terstruktur.

4.4.5 Penghapusan Data Duplikat *(Duplicate Removal)*

Setelah tahap normalisasi, tahap selanjutnya adalah penghapusan data duplikat yang berfungsi untuk menghilangkan data yang memiliki kesamaan atau identik. Tahap ini penting untuk mencegah bias dalam proses pelatihan model akibat pengulangan data yang dapat memengaruhi representasi data secara tidak seimbang. Hasil dari proses penghapusan data duplikat disajikan pada Tabel 4.9.

Tabel 4. 9 Hasil Penghapusan Duplikat

Deskripsi	Jumlah Data
Jumlah data sebelum penghapusan	10.834
Jumlah data setelah penghapusan	10.398
Jumlah duplikasi yang dihapus	436

Berdasarkan Tabel 4.9, proses *duplicate removal* berhasil mengurangi sebanyak 436 data yang teridentifikasi sebagai duplikasi. Hal ini menunjukkan bahwa dataset awal masih mengandung data redundan yang berpotensi memengaruhi kualitas analisis. Dengan dihilangkannya data duplikat, dataset menjadi lebih representatif dan mampu meningkatkan kualitas serta keandalan model dalam proses klasifikasi sentimen.

4.4.6 *Filtering* Topik

Setelah tahap penghapusan data duplikat, tahap selanjutnya adalah *filtering* topik yang bertujuan untuk memastikan bahwa data yang digunakan benar-benar relevan dengan fokus penelitian. Proses ini dilakukan dengan pendekatan berbasis *keyword matching* menggunakan daftar kata kunci yang berkaitan dengan kecerdasan buatan (AI) dan konteks pendidikan.

Berdasarkan hasil implementasi *filtering*, jumlah data sebelum penyaringan adalah sebanyak 10.398 komentar. Setelah dilakukan proses *filtering* berdasarkan kesesuaian kata kunci tematik, diperoleh sebanyak 4.726 komentar yang relevan. Dengan demikian, sebanyak 5.672 data dieliminasi karena tidak memiliki keterkaitan dengan topik penelitian.

Hasil dari proses *filtering* topik disajikan pada Tabel 4.10.

Tabel 4. 10 Hasil *Filtering*

Deskripsi	Jumlah Data
Sebelum	10.834
Sesudah	4.726
Terhapus	5.672

Berdasarkan Tabel 4.10, terjadi pengurangan jumlah data yang signifikan setelah proses *filtering*. Hal ini menunjukkan bahwa sebagian besar data awal tidak relevan dengan topik penelitian. Oleh karena itu, tahap *filtering* berperan penting dalam meningkatkan kualitas dataset dengan memastikan bahwa hanya data yang sesuai konteks yang digunakan dalam proses analisis. Dengan demikian, dataset yang dihasilkan menjadi lebih terarah, bersih, dan relevan untuk digunakan dalam tahap pemodelan selanjutnya.

4.4.7 Representasi Input dan Tokenisasi

Setelah tahap *pre-processing* selesai dilakukan, data teks hasil normalisasi kemudian direpresentasikan ke dalam bentuk numerik agar dapat diproses oleh model berbasis transformer, yaitu BERT dan IndoRoBERTa. Tahap ini disebut tokenisasi, yaitu proses mengubah teks menjadi serangkaian token yang memiliki indeks numerik (token ID) berdasarkan *vocabulary* masing-masing model. Selain token utama, setiap input juga dilengkapi dengan special token, *attention mask*, dan *padding* agar struktur kalimat dapat dikenali oleh model secara optimal. Perbedaan arsitektur *tokenizer* pada BERT dan IndoRoBERTa menyebabkan hasil representasi input yang berbeda, sehingga perlu dianalisis secara terpisah.

a. Representasi Input dan Tokenisasi pada BERT

Model BERT menggunakan metode *WordPiece Tokenizer* dalam proses tokenisasi. Metode ini bekerja dengan memecah kata menjadi unit *subword* apabila kata tersebut tidak ditemukan secara utuh dalam *vocabulary* model. Pendekatan ini memungkinkan model untuk tetap mengenali kata baru atau kata tidak baku yang sering muncul pada komentar YouTube (Song *et al*., 2021).

Pada proses representasi input, setiap kalimat diawali dengan token khusus [CLS] yang berfungsi sebagai representasi keseluruhan kalimat untuk tugas klasifikasi, serta diakhiri dengan token [SEP] sebagai penanda akhir kalimat. Selain itu, digunakan *padding token* bernilai 0 untuk menyeragamkan panjang seluruh input sehingga dapat diproses secara batch oleh model.

Contoh hasil tokenisasi pada model BERT dapat dilihat pada Tabel 4.11.

Tabel 4. 11 Representasi Token ID Data *Normalized* pada BERT

Normalized	Token ID BERT
intinya rumus membantu manusia hanya data saat informasi kurang lengkap ai berbahaya karena ai haus informasi program data menjadi prioritas ai hanya algoritma	[2, 7004, 6152, 1055, 666, 344, 1006, 305, 683, 1057, 1556, 8300, 4199, 211, 8300, 13524, 683, 986, 1006, 234, 7022, 8300, 344, 3]
tinggalkan jejak di sini semoga pekerjaan sales tetap aman dari ai	[2, 8966, 9081, 26, 1585, 1671, 1367, 8612, 830, 1703, 98, 8300, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
ai merupakan penemuan terbesar manusia pengkodean berkembang menjadi inteligensia dan super inteligensia sangat luar biasa	[2, 8300, 407, 6847, 2805, 666, 8487, 791, 5, 2016, 234, 4430, 27291, 13525, 41, 2721, 4430, 27291, 13525, 310, 892, 1167, 3, 0]
cara kerja ai dapat menggambar dan menjawab berbagai pertanyaan	[2, 354, 494, 8300, 173, 10584, 41, 3199, 628, 1544, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
seberapa canggih ai tidak dapat menembus ruang waktu	[2, 6328, 6149, 8300, 119, 173, 8994, 1516, 486, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

Berdasarkan Tabel 4.10, model BERT menghasilkan representasi numerik berupa token ID yang diawali dengan token khusus bernilai 2 sebagai penanda awal kalimat ([CLS]) dan diakhiri dengan token bernilai 3 sebagai penanda akhir kalimat ([SEP]). Setelah token akhir, sistem menambahkan nilai 0 sebagai *padding* token untuk menyeragamkan panjang seluruh input menjadi 24 token. Hal ini diperlukan agar seluruh data dapat diproses secara seragam dalam proses pelatihan model.

BERT menggunakan metode *WordPiece Tokenizer* yang memungkinkan pemecahan kata menjadi unit *subword* apabila suatu kata tidak tersedia secara utuh dalam *vocabulary* model. Pendekatan ini membantu model memahami berbagai variasi kata, termasuk istilah baru dan kata tidak baku yang sering muncul pada komentar YouTube.

b. Representasi Input dan Tokenisasi pada IndoRoBERTa

Berbeda dengan BERT, model IndoRoBERTa menggunakan metode *Byte-Pair Encoding* (BPE) dalam proses tokenisasi. Metode ini cenderung mempertahankan bentuk kata secara lebih utuh dibandingkan *WordPiece*, sehingga lebih adaptif terhadap karakteristik bahasa Indonesia, khususnya pada teks informal yang banyak ditemukan dalam komentar media sosial (Gutierrez-Vasques *et al*., 2021).

Pada proses representasi input, IndoRoBERTa menggunakan token khusus <s> sebagai penanda awal kalimat dan </s> sebagai penanda akhir kalimat. *Padding* pada model ini menggunakan nilai token 1 sesuai dengan konfigurasi tokenizer bawaan model. Selain itu, *attention mask* juga digunakan untuk membedakan token aktual dengan token *padding* agar model dapat fokus pada informasi yang relevan.

Contoh hasil tokenisasi dari data *normalized* pada model IndoRoBERTa dapat dilihat pada Tabel 4.12.

Tabel 4. 12 Representasi Token ID Data *Normalized* pada IndoRoBERTa

Data *Normalized*	Token ID IndoRoBERTa
intinya rumus membantu manusia hanya data saat informasi kurang lengkap ai berbahaya karena ai haus informasi program data menjadi prioritas ai hanya algoritma	[0, 507, 524, 11657, 2171, 1436, 835, 2586, 660, 2671, 1769, 4165, 44601, 7680, 594, 44601, 30437, 2671, 1824, 2586, 446, 15153, 44601, 2]
tinggalkan jejak di sini semoga pekerjaan sales tetap aman dari ai	[0, 88, 801, 280, 9896, 277, 4420, 35528, 3712, 269, 1832, 645, 5344, 323, 44601, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1]
ai merupakan penemuan terbesar manusia pengkodean berkembang menjadi inteligensia dan super inteligensia sangat luar biasa	[0, 431, 432, 6747, 1843, 1436, 49068, 2289, 446, 28330, 862, 28711, 291, 5664, 28330, 862, 28711, 874, 1423, 1937, 2, 1, 1, 1]
cara kerja ai dapat menggambar dan	[0, 505, 2388, 44601, 513, 16015, 291, 7454,

menjawab berbagai pertanyaan	1088, 7998, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
seberapa canggih ai tidak dapat menembus ruang waktu	[0, 87, 593, 13275, 44601, 463, 513, 10769, 2403, 1074, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

Berdasarkan Tabel 4.12, IndoRoBERTa menghasilkan *token ID* yang diawali dengan nilai 0 sebagai token awal (<s>) dan diakhiri dengan nilai 2 sebagai token akhir (</s>). Setelah itu, digunakan nilai 1 sebagai *padding token* untuk menyeragamkan panjang input. Berbeda dengan BERT, IndoRoBERTa menggunakan metode *Byte-Pair Encoding* (BPE) yang cenderung mempertahankan bentuk kata secara lebih utuh.

Karakteristik ini membuat IndoRoBERTa lebih adaptif terhadap bahasa Indonesia, khususnya pada komentar YouTube yang mengandung bahasa informal, singkatan, dan istilah modern seperti kecerdasan buatan (*Artificial Intelligence*). Perbedaan representasi token ini menjadi salah satu faktor yang memengaruhi performa klasifikasi sentimen pada kedua model.

c. Perbandingan Tokenisasi BERT dan IndoRoBERTa

Perbedaan metode tokenisasi antara BERT dan IndoRoBERTa menghasilkan representasi input yang berbeda terhadap data komentar yang sama. BERT dengan *WordPiece Tokenizer* cenderung memecah kata menjadi beberapa unit *subword* apabila kata tersebut tidak tersedia secara utuh dalam *vocabulary*, sedangkan IndoRoBERTa dengan *Byte-Pair Encoding* lebih mampu mempertahankan bentuk kata secara lengkap (Venkatesan & Arulanand, 2022).

Perbedaan ini menunjukkan bahwa IndoRoBERTa lebih adaptif terhadap kosakata modern dan bahasa informal yang banyak ditemukan pada komentar YouTube, sedangkan BERT tetap unggul dalam representasi kontekstual yang luas karena proses *pre-training* global yang dimilikinya. Pada penelitian ini, karakteristik data komentar publik yang bersifat informal dan dinamis menjadi salah satu alasan penting dalam membandingkan kedua model (Venkatesan & Arulanand, 2022; Choo & Kim, 2023).

Dengan demikian, proses tokenisasi menjadi tahap yang sangat penting karena berpengaruh langsung terhadap kemampuan model dalam memahami konteks kalimat. Perbedaan representasi input ini selanjutnya berkontribusi

terhadap hasil performa klasifikasi sentimen yang diperoleh pada tahap evaluasi model.

4.5 Hasil Pengujian dan Analisis Model Klasifikasi Sentimen

Hasil pengujian model klasifikasi sentimen dilakukan pada tiga skenario pembagian data (70:15:15, 80:10:10, dan 90:5:5) untuk menganalisis pengaruh proporsi data terhadap kinerja BERT dan IndoRoBERTa. Pengujian mencakup tahap *baseline* dan *fine-tuning* dengan *class weight* guna mengatasi ketidakseimbangan kelas. Evaluasi menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil kemudian disajikan melalui analisis per skenario, perbandingan antar skenario, serta *confusion matrix* untuk mengidentifikasi kesalahan klasifikasi.

4.5.1 Skenario 1 (70:15:15)

Skenario pertama menggunakan pembagian dataset sebesar 70% data latih, 15% data validasi, dan 15% data uji, dengan rincian 3308 data latih, 709 data validasi, dan 709 data uji. Skenario ini bertujuan untuk mengevaluasi kinerja model dengan proporsi data latih yang lebih dominan agar model dapat mempelajari pola data secara lebih optimal. Hasil pengujian pada skenario ini meliputi tahap *baseline*, proses pelatihan, dan evaluasi akhir.

a. Hasil *Baseline*

Hasil pengujian awal *(baseline)* disajikan pada Tabel 4.13.

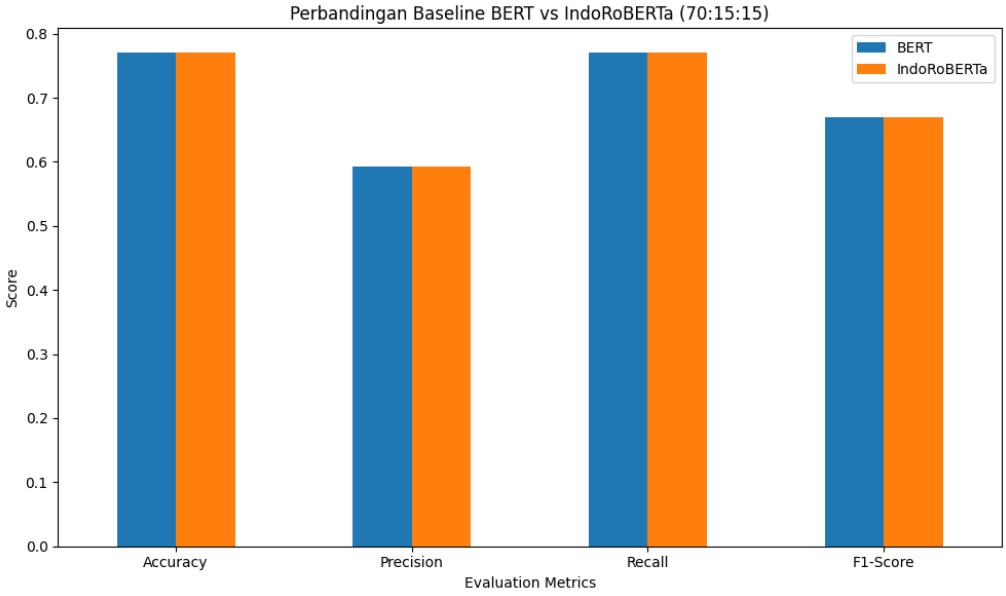
Tabel 4. 13 Hasil Pengujian *Baseline* Model BERT dan IndoRoBERTa

Model	Accuracy	Precision	Recall	F1-Score
BERT	0.77	0.59	0.77	0.67
IndoRoBERTa	0.77	0.59	0.77	0.67

Berdasarkan Tabel 4.13, kedua model menunjukkan performa yang sama dengan nilai *accuracy* sebesar 0,77, *precision* sebesar 0,59, *recall* sebesar 0,77, dan *f1-score* sebesar 0,67. Hasil ini menunjukkan bahwa pada tahap baseline, BERT dan IndoRoBERTa belum memiliki perbedaan performa yang signifikan. Nilai *recall* sebesar 0,77 menunjukkan bahwa model mampu mengenali sebagian

besar data dengan cukup baik, namun *precision* dan *f1-score* yang masih relatif rendah menunjukkan bahwa kemampuan klasifikasi belum merata pada seluruh kelas sentimen. Kondisi ini mengindikasikan adanya ketidakseimbangan kelas *(class imbalance)*, di mana distribusi data yang tidak merata menyebabkan model menjadi bias terhadap kelas mayoritas. Oleh karena itu, diperlukan proses optimasi lanjutan untuk meningkatkan performa klasifikasi pada seluruh kelas sentimen.

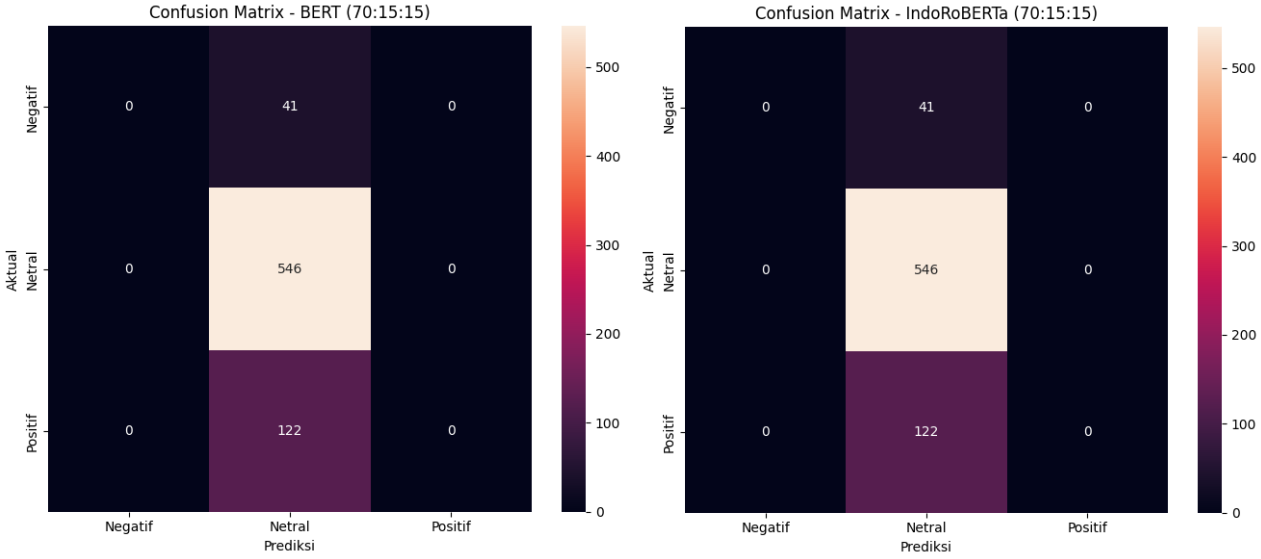
Untuk memperjelas perbandingan performa kedua model, hasil *baseline* juga divisualisasikan dalam bentuk grafik sebagaimana disajikan pada Gambar 4.1.



Gambar 4. 1 Perbandingan *Baseline* BERT dan IndoRoBERTa (70:15:15)

Berdasarkan Gambar 4.1, terlihat bahwa kedua model memiliki nilai evaluasi yang identik pada seluruh metrik. *Accuracy* dan *recall* sama-sama mencapai 0,77, precision sebesar 0,59, serta f1-score sebesar 0,67. Hasil ini menegaskan bahwa pada tahap awal, kedua model masih belum optimal dalam menangani klasifikasi multikelas, terutama pada kelas negatif dan positif yang jumlah datanya lebih sedikit. Dengan demikian, tahap *baseline* menunjukkan bahwa diperlukan optimasi lebih lanjut melalui *preprocessing*, penyesuaian *hyperparameter*, dan penanganan data tidak seimbang agar performa model dapat meningkat secara lebih merata pada seluruh kelas sentimen.

Untuk melihat distribusi prediksi benar dan salah secara rinci, hasil evaluasi *baseline* BERT dan IndoRoBERTa disajikan dalam satu *confusion matrix* gabungan pada Gambar 4.2. Visualisasi ini menampilkan pola klasifikasi yang tidak tercermin langsung pada nilai *accuracy*, *precision*, *recall*, dan *f1-score*.



Gambar 4. 2 *Confusion Matrix Baseline* BERT dan IndoRoBERTa (70:15:15)

Berdasarkan Gambar 4.2, kedua model menunjukkan pola yang sama, yaitu seluruh data uji diprediksi sebagai kelas netral. Dari 709 data, sebanyak 546 data netral terklasifikasi benar, sedangkan 41 data negatif dan 122 data positif seluruhnya salah diprediksi sebagai netral. Tidak ada prediksi pada kelas negatif maupun positif. Hal ini menandakan bahwa model hanya mengenali kelas mayoritas dan mengalami bias tinggi akibat *class imbalance*. Dampaknya, nilai *precision*, *recall*, dan *f1-score* pada kelas negatif dan positif bernilai 0,00. Secara keseluruhan, kondisi ini menunjukkan perlunya optimasi lanjutan agar model mampu mengenali seluruh kelas secara seimbang.

Temuan ini menunjukkan bahwa meskipun BERT dan IndoRoBERTa merupakan model berbasis transformer yang memiliki kemampuan representasi bahasa yang kuat, performa model tetap dipengaruhi oleh distribusi data yang digunakan. Dominasi kelas netral menyebabkan model lebih sering memilih kelas mayoritas untuk meminimalkan kesalahan prediksi secara keseluruhan. Akibatnya, karakteristik linguistik pada kelas negatif dan positif belum dapat

dipelajari secara optimal. Kondisi ini mengindikasikan bahwa kualitas distribusi data memiliki peran penting dalam membangun model klasifikasi sentimen yang mampu mengenali seluruh kelas secara seimbang.

b. *Fine Tuning*

Pada tahap *fine-tuning*, model BERT dan IndoRoBERTa dilatih menggunakan data hasil pembagian skenario penelitian untuk meningkatkan performa klasifikasi sentimen dibandingkan tahap *baseline*. Proses pelatihan dilakukan selama 5 *epoch* dengan penerapan *class weight* untuk mengatasi ketidakseimbangan kelas pada dataset. Evaluasi model dilakukan menggunakan metrik *training loss*, *validation loss*, *accuracy*, *precision*, *recall*, dan *F1-score* untuk mengamati perkembangan performa pada setiap iterasi pelatihan.

Hasil training model BERT pada setiap *epoch* disajikan pada Tabel 4.14.

Tabel 4. 14 Hasil *Training* BERT per Epoch

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1-Score
1	0.937994	1.284401	0.808181	0.758178	0.808181	0.782065
2	0.802741	0.590482	0.761636	0.868800	0.761636	0.797248
3	0.594332	0.731680	0.861777	0.863297	0.861777	0.860136
4	0.407291	0.742643	0.874471	0.878043	0.874471	0.875605
5	0.289562	0.768970	0.887165	0.889273	0.887165	0.888047

Berdasarkan Tabel 4.14, model BERT menunjukkan peningkatan performa yang konsisten selama proses pelatihan. Nilai *training loss* menurun dari 0.937994 pada *epoch* pertama menjadi 0.289562 pada *epoch* kelima, yang menandakan bahwa model semakin baik dalam mempelajari pola data latih. Akurasi model juga meningkat dari 0.808181 menjadi 0.887165, disertai peningkatan nilai *precision* dari 0.758178 menjadi 0.889273, *recall* dari 0.808181 menjadi 0.887165, serta *F1-score* dari 0.782065 menjadi 0.888047.

Meskipun pada *epoch* kedua terjadi penurunan akurasi menjadi 0.761636, performa model kembali meningkat secara stabil pada *epoch* berikutnya hingga mencapai hasil terbaik pada *epoch* kelima. Hal ini menunjukkan bahwa model BERT mampu beradaptasi dengan baik terhadap distribusi data setelah proses *fine-tuning* dilakukan.

Selanjutnya, hasil *training* model IndoRoBERTa ditunjukkan pada Tabel 4.15.

Tabel 4. 15 Hasil *Training* IndoRoBERTa per *Epoch*

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1-Score
1	0.810635	0.810642	0.867419	0.882375	0.867419	0.860380
2	0.474906	0.380375	0.930889	0.934789	0.930889	0.932146
3	0.248355	0.431228	0.957687	0.957412	0.957687	0.957383
4	0.108640	0.551791	0.957687	0.957255	0.957687	0.957198
5	0.041414	0.592157	0.954866	0.954415	0.954866	0.954368

Berdasarkan Tabel 4.14, model IndoRoBERTa menunjukkan peningkatan performa yang cepat selama proses *fine-tuning*. Nilai *training loss* menurun dari 0.810635 pada *epoch* pertama menjadi 0.041414 pada *epoch* kelima, yang menunjukkan bahwa model semakin baik dalam mempelajari pola data latih. Akurasi meningkat dari 0.867419 pada *epoch* pertama dan mencapai nilai tertinggi sebesar 0.957687 pada *epoch* ketiga dan keempat. Nilai *precision, recall,* dan *F1-score* juga mengalami peningkatan yang signifikan.

Performa terbaik dicapai pada *epoch* ketiga dan keempat dengan *accuracy* sebesar 0.957687 dan *F1-score* masing-masing sebesar 0.957383 dan 0.957198. Pada *epoch* kelima terjadi sedikit penurunan akurasi menjadi 0.954866 serta peningkatan validation loss menjadi 0.592157, yang menunjukkan gejala *overfitting* ringan.

Performa IndoRoBERTa yang lebih tinggi dibandingkan BERT menunjukkan bahwa model tersebut memiliki kemampuan yang lebih baik dalam memahami konteks bahasa Indonesia. Hal ini dapat dipengaruhi oleh proses *pre-training* yang menggunakan korpus bahasa Indonesia dalam jumlah besar sehingga representasi semantik yang dihasilkan lebih sesuai dengan karakteristik komentar YouTube yang didominasi bahasa informal, singkatan, dan variasi kosakata. Dengan demikian, IndoRoBERTa mampu melakukan generalisasi yang lebih baik terhadap data yang tidak ditemukan selama proses pelatihan.

c. Hasil Evaluasi Akhir

Hasil evaluasi akhir model setelah proses *fine-tuning* disajikan pada Tabel 4.16.

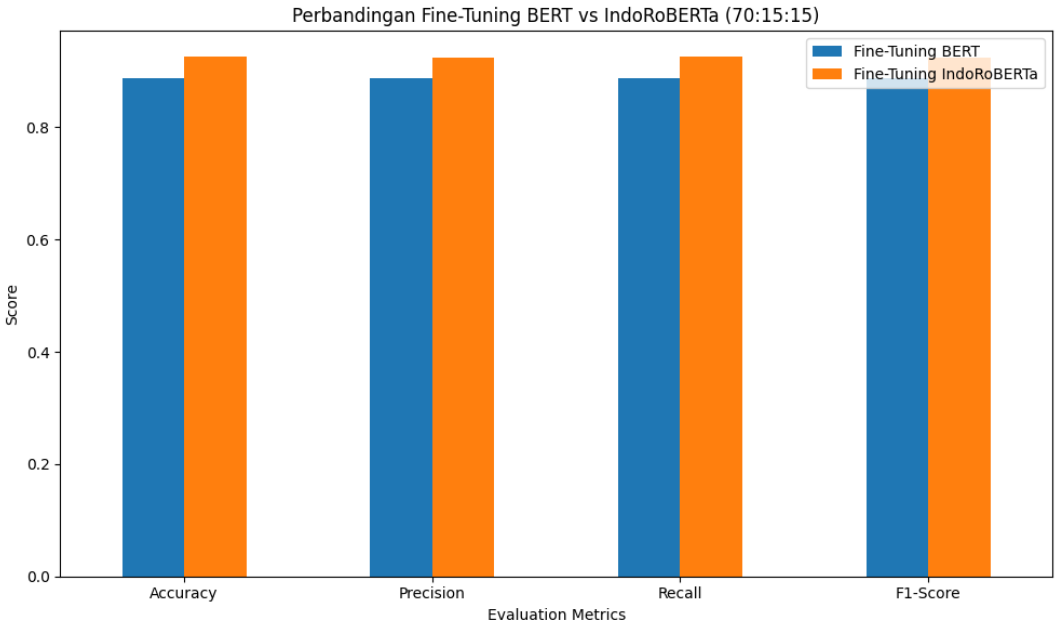
Tabel 4. 16 Hasil Evaluasi Skenario 70:15:15

Model	Accuracy	Precision	Recall	F1-Score
Fine-Tuning BERT	0.89	0.89	0.89	0.89
Fine-Tuning IndoRoBERTa	0.93	0.92	0.93	0.92

Berdasarkan Tabel 4.16, hasil evaluasi akhir menunjukkan bahwa kedua model mengalami peningkatan performa yang signifikan dibandingkan tahap *baseline*. Model BERT memperoleh *accuracy* sebesar 0.89 dengan *precision*, *recall*, dan *f1-score* masing-masing sebesar 0.89. Sementara itu, model IndoRoBERTa menunjukkan performa yang lebih baik dengan *accuracy* sebesar 0.93, *precision* sebesar 0.92, *recall* sebesar 0.93, dan *f1-score* sebesar 0.92.

Hasil tersebut menunjukkan bahwa proses *fine-tuning* berhasil meningkatkan kemampuan model dalam mengklasifikasikan sentimen komentar secara lebih akurat pada seluruh kelas. IndoRoBERTa unggul dibandingkan BERT karena mampu menghasilkan performa yang lebih tinggi dan lebih seimbang, terutama dalam mengenali kelas negatif dan positif yang sebelumnya sulit diklasifikasikan pada tahap *baseline*.

Untuk memperjelas perbandingan performa akhir antara model BERT dan IndoRoBERTa setelah proses *fine-tuning*, hasil evaluasi divisualisasikan dalam bentuk grafik pada Gambar 4.3. Grafik ini menampilkan perbandingan nilai *accuracy*, *precision*, *recall*, dan *f1-score* sehingga perbedaan performa kedua model dapat terlihat secara lebih jelas.

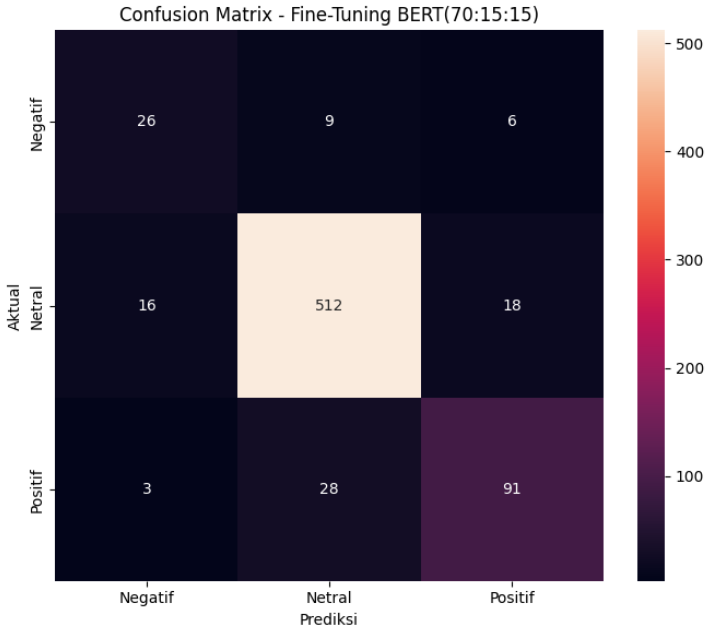


Gambar 4. 3 Perbandingan *Fine-Tuning* BERT vs IndoRoBERTa (70:15:15)

Berdasarkan Gambar 4.3, model IndoRoBERTa menunjukkan performa yang lebih baik dibandingkan BERT pada seluruh metrik evaluasi. Model BERT memperoleh *accuracy*, *precision*, *recall*, dan *f1-score* masing-masing sebesar 0,89. Sementara itu, IndoRoBERTa memperoleh *accuracy* sebesar 0,93, *precision* sebesar 0,92, *recall* sebesar 0,93, dan f1-score sebesar 0,92.

Selisih nilai tersebut menunjukkan bahwa IndoRoBERTa lebih mampu melakukan klasifikasi sentimen secara akurat dan seimbang pada seluruh kelas, terutama pada kelas negatif dan positif yang sebelumnya sulit dikenali pada tahap *baseline*. Hasil ini menegaskan bahwa proses *fine-tuning* berhasil meningkatkan performa kedua model, dengan IndoRoBERTa sebagai model yang paling optimal pada skenario 70:15:15.

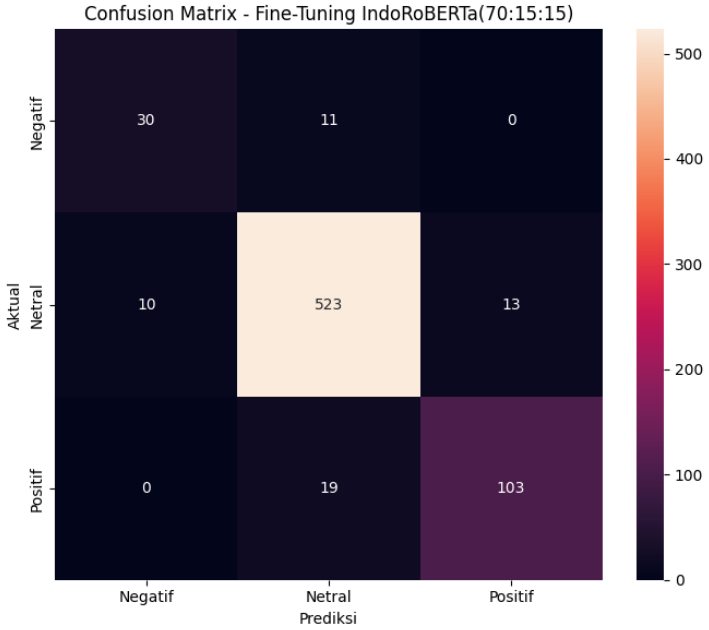
Untuk memberikan gambaran yang lebih rinci terkait performa model, hasil klasifikasi BERT setelah proses *fine-tuning* divisualisasikan menggunakan *confusion matrix* pada Gambar 4.4. Visualisasi ini menunjukkan distribusi prediksi model terhadap masing-masing kelas aktual, yaitu negatif, netral, dan positif.



Gambar 4. 4 *Confusion Matrix Fine-Tuning* BERT (70:15:15)

Berdasarkan Gambar 4.4, model BERT mampu mengklasifikasikan kelas netral dengan sangat baik, ditunjukkan oleh nilai prediksi benar sebanyak 512 data. Namun, masih terdapat kesalahan klasifikasi pada kelas negatif yang diprediksi sebagai netral (9 data) dan positif (6 data). Selain itu, pada kelas positif terdapat 28 data yang keliru diklasifikasikan sebagai netral dan 3 data sebagai negatif. Hal ini menunjukkan bahwa meskipun performa keseluruhan model cukup tinggi, BERT masih mengalami kesulitan dalam membedakan secara tegas antara kelas netral dengan kelas lainnya, terutama pada data dengan ambiguitas sentimen.

Selanjutnya, untuk membandingkan secara lebih mendalam, *confusion matrix* hasil *fine-tuning* IndoRoBERTa ditampilkan pada Gambar 4.5. Visualisasi ini memberikan informasi distribusi prediksi yang lebih presisi pada masing-masing kelas dibandingkan model sebelumnya.



Gambar 4. 5 *Confusion Matrix Fine-Tuning* IndoRoBERTa (70:15:15)

Berdasarkan Gambar 4.5, IndoRoBERTa menunjukkan peningkatan performa yang signifikan. Model ini mampu mengklasifikasikan kelas netral dengan sangat akurat, dengan 523 prediksi benar. Pada kelas negatif, jumlah prediksi benar mencapai 30 data dengan kesalahan yang relatif lebih sedikit dibandingkan BERT. Selain itu, pada kelas positif, model berhasil mengklasifikasikan 103 data dengan benar, dengan hanya sebagian kecil yang salah diprediksi sebagai netral (19 data). Tidak terdapat kesalahan klasifikasi antara kelas negatif dan positif secara langsung, yang menunjukkan bahwa model memiliki kemampuan pemisahan kelas yang lebih baik.

Secara keseluruhan, *confusion matrix* IndoRoBERTa memperlihatkan distribusi kesalahan yang lebih kecil dan lebih terfokus dibandingkan BERT, sehingga memperkuat hasil evaluasi sebelumnya bahwa IndoRoBERTa memiliki performa yang lebih optimal dalam klasifikasi sentimen pada skenario pembagian data 70:15:15.

4.5.2 Skenario 2 (80:10:10)

Skenario kedua menggunakan pembagian dataset sebesar 80% data latih, 10% data validasi, dan 10% data uji, dengan rincian 3780 data latih, 473 data

validasi, dan 473 data uji. Proporsi ini memberikan porsi data pelatihan yang lebih besar dibandingkan skenario sebelumnya, sehingga diharapkan model dapat mempelajari pola data secara lebih optimal. Hasil pengujian pada skenario ini disajikan secara bertahap meliputi *baseline*, proses pelatihan, dan evaluasi akhir.

a. Hasil *Baseline*

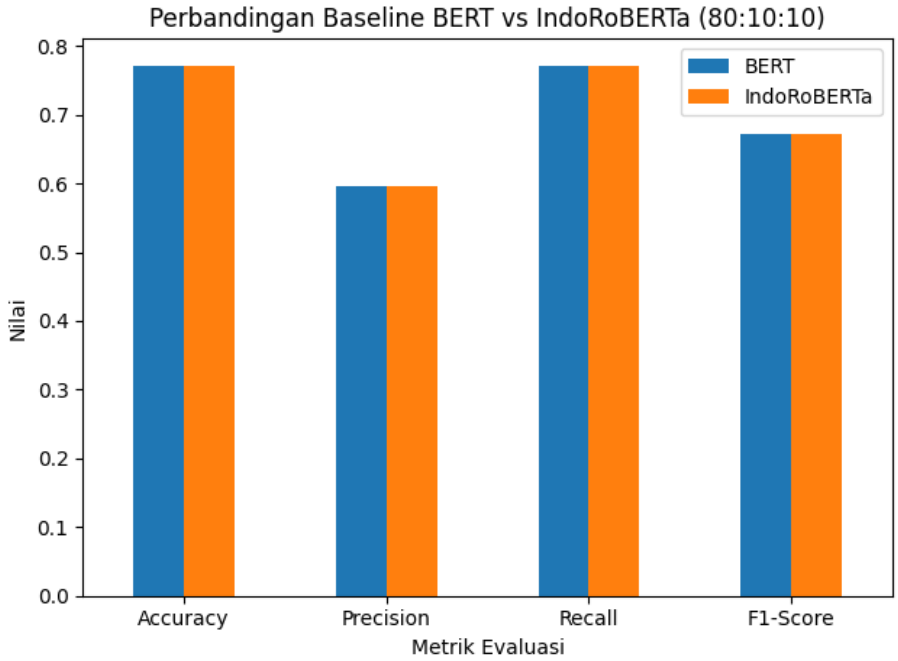
Hasil pengujian awal *(baseline)* pada skenario ini disajikan pada Tabel 4.17.

Tabel 4. 17 *Baseline* Skenario 80:10:10

Model	Accuracy	Precision	Recall	F1-Score
BERT	0.77	0.60	0.77	0.67
IndoRoBERTa	0.77	0.60	0.77	0.67

Berdasarkan Tabel 4.17, model BERT dan IndoRoBERTa menunjukkan hasil *baseline* yang sama pada skenario 80:10:10 dengan nilai accuracy sebesar 0,77, *precision* sebesar 0,60, *recall* sebesar 0,77, dan *f1-score* sebesar 0,67. Hasil ini menunjukkan bahwa kedua model memiliki performa awal yang belum optimal karena masih cenderung memprediksi pada kelas mayoritas. Nilai *precision* dan *f1-score* yang masih relatif rendah menunjukkan bahwa kemampuan model dalam membedakan seluruh kelas sentimen belum merata. Kondisi ini mengindikasikan adanya ketidakseimbangan kelas *(class imbalance)*, sehingga diperlukan proses *fine-tuning* untuk meningkatkan performa klasifikasi secara lebih akurat pada seluruh kelas sentimen.

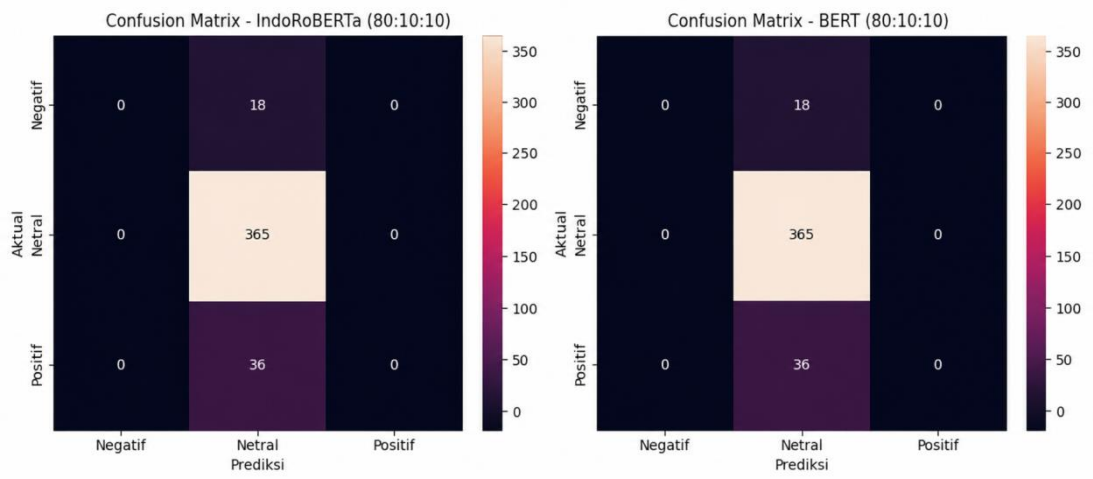
Untuk memberikan gambaran yang lebih jelas mengenai perbandingan performa kedua model, hasil *baseline* juga ditampilkan dalam bentuk grafik pada Gambar 4.6.



Gambar 4. 6 Perbandingan *Baseline* BERT vs IndoRoBERTa (80:10:10)

Berdasarkan Gambar 4.6, kedua model menunjukkan hasil evaluasi yang sama pada seluruh metrik pengukuran. Nilai *accuracy* dan *recall* masing-masing sebesar 0,77, *precision* sebesar 0,59, serta *f1-score* sebesar 0,67. Hasil ini menunjukkan bahwa pada tahap *baseline*, performa BERT dan IndoRoBERTa masih belum optimal dalam melakukan klasifikasi multikelas, khususnya pada kelas dengan jumlah data yang lebih sedikit. Oleh karena itu, diperlukan optimasi lanjutan seperti penanganan ketidakseimbangan data agar performa klasifikasi dapat meningkat secara lebih merata pada seluruh kelas sentimen.

Untuk melihat distribusi prediksi benar dan salah pada setiap kelas sentimen secara lebih rinci, hasil evaluasi *baseline* model BERT dan IndoRoBERTa juga divisualisasikan menggunakan *confusion matrix*. Visualisasi ini membantu menunjukkan pola klasifikasi model serta kecenderungan prediksi pada masing-masing kelas yang tidak terlihat secara langsung dari nilai *accuracy*, *precision*, *recall*, dan *f1-score*, sebagaimana disajikan pada Gambar 4.7.



Gambar 4. 7 *Confusion Matrix Baseline* BERT dan Indoroberta (80:1.0:10)

Berdasarkan Gambar 4.7, terlihat bahwa model BERT dan IndoRoBERTa memiliki pola prediksi yang identik, yaitu seluruh data uji diprediksi ke dalam kelas netral. Sebanyak 365 data netral berhasil diprediksi dengan benar sebagai netral, sedangkan seluruh data dari kelas lainnya juga diprediksi sebagai netral. Tidak terdapat prediksi pada kelas negatif maupun positif. Kondisi ini menunjukkan bahwa kedua model masih sangat bergantung pada kelas mayoritas dan belum mampu membedakan kelas sentimen secara seimbang. Hal ini memperkuat bahwa pada tahap *baseline* masih terjadi *class imbalance*, sehingga diperlukan proses *fine-tuning* untuk meningkatkan kemampuan klasifikasi pada seluruh kelas sentimen.

b. *Fine Tuning*

Setelah tahap *baseline* dilakukan, proses berikutnya adalah *fine-tuning* model BERT dan IndoRoBERTa menggunakan data latih dengan skenario pembagian 80:10:10 serta penerapan *class weight* untuk mengatasi ketidakseimbangan kelas. Pada tahap ini, model dilatih selama 5 *epoch* untuk mengamati perkembangan performa pada setiap iterasi pelatihan. Evaluasi dilakukan menggunakan metrik *training loss*, *validation loss*, *accuracy*, *precision*, *recall*, dan *F1-score*.

Hasil pelatihan per *epoch* pada model BERT dapat dilihat pada Tabel 4.18.

Tabel 4. 18 Hasil *Training* BERT per *Epoch*

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1-Score
1	0.902200	0.720342	0.849894	0.836543	0.849894	0.834760
2	0.506803	0.540699	0.883721	0.911584	0.883721	0.891955
3	0.320928	0.490759	0.934461	0.933929	0.934461	0.934143
4	0.171905	0.521507	0.940803	0.940195	0.940803	0.940384
5	0.105865	0.549738	0.942918	0.942510	0.942918	0.942444

Berdasarkan Tabel 4.18, *training loss* BERT menurun signifikan dari 0.902200 menjadi 0.105865, disertai peningkatan *accuracy* dari 0.849894 menjadi 0.942918. Nilai *precision*, *recall*, dan *F1-score* juga meningkat secara konsisten, menunjukkan kestabilan performa. Meskipun *validation loss* sedikit meningkat pada *epoch* akhir, performa evaluasi tetap membaik, yang mengindikasikan kemampuan generalisasi yang cukup baik serta kontribusi *class weight* dalam menangani kelas minoritas.

Hasil training model IndoRoBERTa pada setiap epoch disajikan pada Tabel 4.19.

Tabel 4. 19 Hasil *Training* IndoRoBERTa per *Epoch*

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1-Score
1	0.764297	0.506881	0.902748	0.909040	0.902748	0.905001
2	0.427884	0.457153	0.932347	0.936016	0.932347	0.933087
3	0.229752	0.517209	0.949260	0.949209	0.949260	0.948489
4	0.127833	0.521703	0.955603	0.955245	0.955603	0.954746
5	0.054804	0.572404	0.947146	0.946881	0.947146	0.946211

Hasil pelatihan IndoRoBERTa pada Tabel 4.19 menunjukkan performa yang lebih tinggi. *Training loss* menurun dari 0.764297 menjadi 0.054804, dengan *accuracy* terbaik sebesar 0.955603 dan *F1-score* 0.954746 pada *epoch* keempat. Pada *epoch* kelima terjadi penurunan *accuracy* menjadi 0.947146 dan peningkatan *validation loss*, yang mengindikasikan *overfitting* ringan, sehingga *epoch* keempat menjadi titik optimal.

Secara keseluruhan, kedua model menunjukkan peningkatan performa selama proses *fine-tuning*, namun IndoRoBERTa memiliki perkembangan yang lebih baik dibandingkan BERT. BERT mencapai performa terbaik dengan *accuracy* sebesar 0.942918 dan *f1-score* sebesar 0.942444 pada *epoch* kelima, sedangkan IndoRoBERTa mencapai *accuracy* sebesar 0.955603 dan *f1-score* sebesar 0.954746 pada *epoch* keempat. Selain itu, IndoRoBERTa juga menunjukkan konvergensi yang lebih cepat dan stabil selama pelatihan. Hal ini menunjukkan bahwa pada tahap *fine-tuning*, IndoRoBERTa memiliki proses pembelajaran yang lebih optimal dibandingkan BERT pada skenario 80:10:10.

c. Hasil Evaluasi Akhir

Hasil evaluasi akhir model pada skenario ini disajikan pada Tabel 4.20.

Tabel 4. 20 Hasil Evaluasi Skenario 80:10:10

Model	Accuracy	Precision	Recall	F1-Score
Fine-Tuning BERT	0.95	0.95	0.95	0.95
Fine-Tuning IndoRoBERTa	0.95	0.95	0.95	0.94

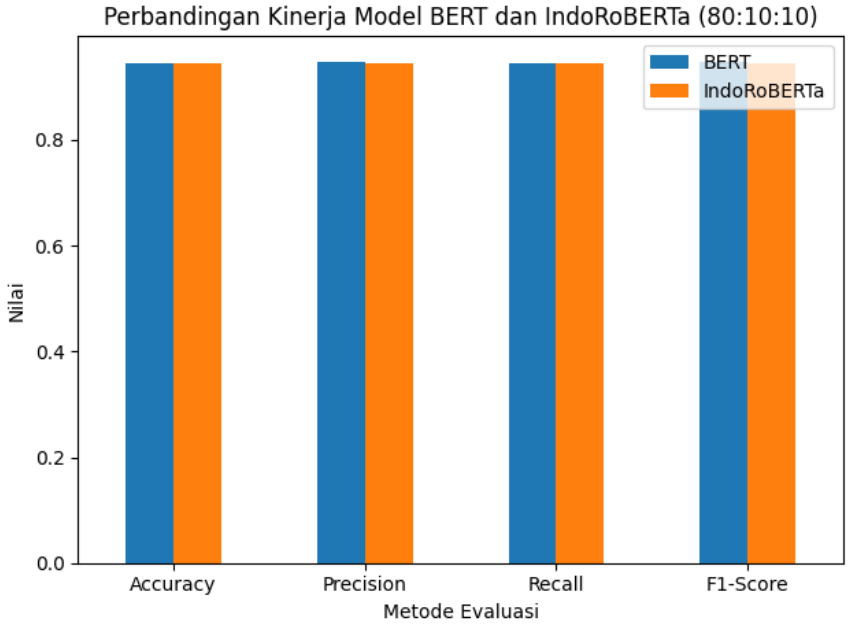
Berdasarkan Tabel 4.20, kedua model menunjukkan performa yang sangat baik dengan nilai metrik yang hampir identik. Model BERT dan IndoRoBERTa sama-sama mencapai akurasi sebesar 0,95, dengan nilai *precision* dan *recall* yang juga seimbang. Hal ini menunjukkan bahwa model tidak hanya mampu meningkatkan akurasi secara keseluruhan, tetapi juga mampu mengklasifikasikan setiap kelas dengan konsisten.

Secara keseluruhan, peningkatan jumlah data latih pada skenario ini memberikan dampak positif terhadap stabilitas dan performa model. Perbedaan kinerja antara BERT dan IndoRoBERTa menjadi semakin kecil, yang mengindikasikan bahwa kedua model telah mampu belajar dengan baik dari data yang tersedia.

Hasil ini menunjukkan bahwa proporsi data latih sebesar 80% memberikan keseimbangan yang baik antara kebutuhan pembelajaran model dan kebutuhan evaluasi. Jumlah data latih yang cukup besar memungkinkan model memperoleh representasi pola sentimen yang lebih lengkap, sementara data validasi dan data

uji masih cukup representatif untuk mengukur kemampuan generalisasi model. Oleh karena itu, skenario 80:10:10 menghasilkan performa yang paling stabil dibandingkan skenario lainnya.

Untuk memperjelas perbandingan performa akhir setelah proses *fine-tuning*, hasil evaluasi model BERT dan IndoRoBERTa divisualisasikan dalam bentuk grafik pada Gambar 4.8. Grafik ini menampilkan nilai *accuracy*, *precision*, *recall*, dan *f1-score* sehingga perbedaan kinerja kedua model terlihat lebih jelas.



Gambar 4. 8 Perbandingan *Fine Tuning* BERT dan IndoRoBERTa (80:10:10)

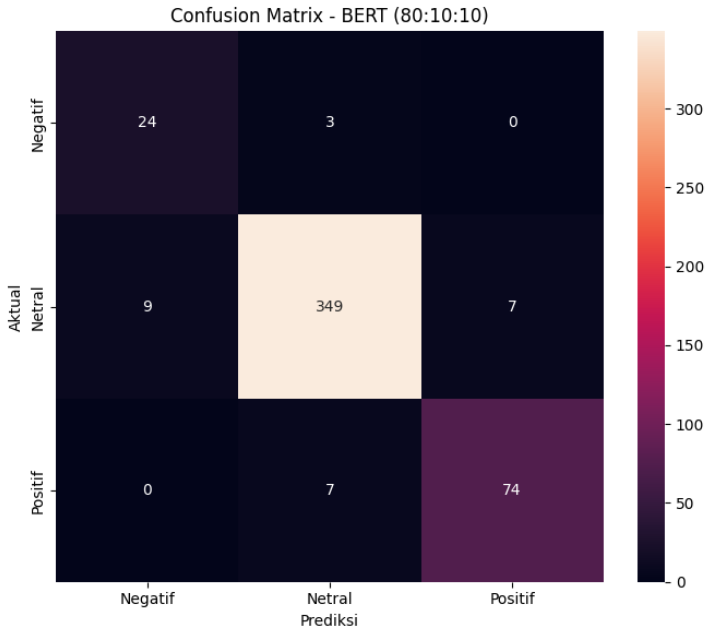
Berdasarkan Gambar 4.8, model BERT dan IndoRoBERTa menunjukkan performa yang hampir setara pada seluruh metrik evaluasi. Model BERT memperoleh *accuracy* sebesar 0,945032, *precision* sebesar 0,948138, *recall* sebesar 0,945032, dan *f1-score* sebesar 0,946072. Sementara itu, IndoRoBERTa memperoleh *accuracy* sebesar 0,945032, *precision* sebesar 0,945071, *recall* sebesar 0,945032, dan *f1-score* sebesar 0,944768.

Selisih nilai yang sangat kecil menunjukkan bahwa kedua model memiliki kemampuan yang relatif sama dalam melakukan klasifikasi sentimen secara akurat dan seimbang pada seluruh kelas. Meskipun demikian, BERT menunjukkan keunggulan tipis pada metrik *precision* dan *f1-score*, yang mengindikasikan

performa yang sedikit lebih baik dalam menjaga keseimbangan antara ketepatan dan kelengkapan prediksi.

Hasil ini menegaskan bahwa pada skenario 80:10:10, kedua model telah mencapai performa optimal dengan perbedaan yang tidak signifikan, sehingga keduanya dapat dianggap sama-sama efektif dalam tugas klasifikasi sentimen.

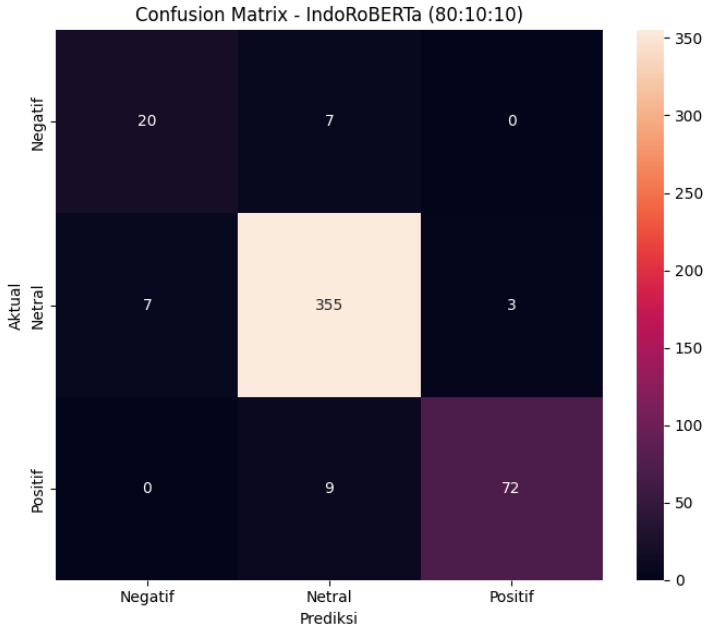
Untuk memperjelas kinerja model, hasil klasifikasi BERT setelah *fine-tuning* ditampilkan dalam bentuk *confusion matrix* pada Gambar 4.9, yang menunjukkan distribusi prediksi terhadap kelas negatif, netral, dan positif.



Gambar 4. 9 *Confusion Matrix Fine-Tuning* BERT (80:15:15)

Berdasarkan Gambar 4.9, model BERT mampu mengklasifikasikan kelas netral dengan sangat baik, ditunjukkan oleh nilai prediksi benar sebanyak 349 data. Namun, masih terdapat kesalahan klasifikasi pada kelas netral yang diprediksi sebagai negatif sebanyak 9 data dan sebagai positif sebanyak 7 data. Selain itu, pada kelas negatif terdapat 3 data yang keliru diklasifikasikan sebagai netral, sedangkan pada kelas positif terdapat 7 data yang salah diprediksi sebagai netral. Hal ini menunjukkan bahwa meskipun performa keseluruhan model cukup tinggi, BERT masih mengalami kesulitan dalam membedakan secara tegas antara kelas netral dengan kelas lainnya, terutama pada data yang memiliki ambiguitas sentimen.

Selanjutnya, untuk analisis perbandingan yang lebih mendalam, *confusion matrix* hasil *fine-tuning* IndoRoBERTa ditampilkan pada Gambar 4.10. Visualisasi ini menyajikan distribusi prediksi model secara lebih detail pada setiap kelas dibandingkan model sebelumnya.



Gambar 4. 10 *Confusion Matrix Fine-Tuning* IndoRoBERTa (80:15:15)

Berdasarkan Gambar 4.10, model IndoRoBERTa mampu mengklasifikasikan kelas netral dengan sangat baik, ditunjukkan oleh nilai prediksi benar sebanyak 355 data. Namun, masih terdapat kesalahan klasifikasi pada kelas netral yang diprediksi sebagai negatif sebanyak 7 data dan sebagai positif sebanyak 3 data. Selain itu, pada kelas negatif terdapat 7 data yang keliru diklasifikasikan sebagai netral, sedangkan pada kelas positif terdapat 9 data yang salah diprediksi sebagai netral. Tidak ditemukan kesalahan klasifikasi langsung antara kelas negatif dan positif. Hal ini menunjukkan bahwa meskipun performa model tergolong tinggi, IndoRoBERTa masih mengalami kesulitan dalam membedakan secara tegas antara kelas netral dengan kelas lainnya, terutama pada data dengan ambiguitas sentimen.

Secara keseluruhan, jika dibandingkan dengan *confusion matrix* BERT, IndoRoBERTa menunjukkan pola kesalahan yang serupa dengan dominasi

kesalahan pada kelas netral, namun dengan distribusi prediksi yang sedikit lebih seimbang pada beberapa kelas.

4.5.3 Skenario 3 (90:5:5)

Skenario ketiga menggunakan pembagian dataset sebesar 90% data latih, 5% data validasi, dan 5% data uji, dengan rincian 4253 data latih, 236 data validasi, dan 237 data uji. Proporsi ini memberikan jumlah data pelatihan yang paling besar dibandingkan skenario sebelumnya, sehingga model memiliki kesempatan yang lebih luas dalam mempelajari pola data. Hasil pengujian pada skenario ini disajikan melalui tahap *baseline*, proses pelatihan, dan evaluasi akhir.

a. Hasil Baseline

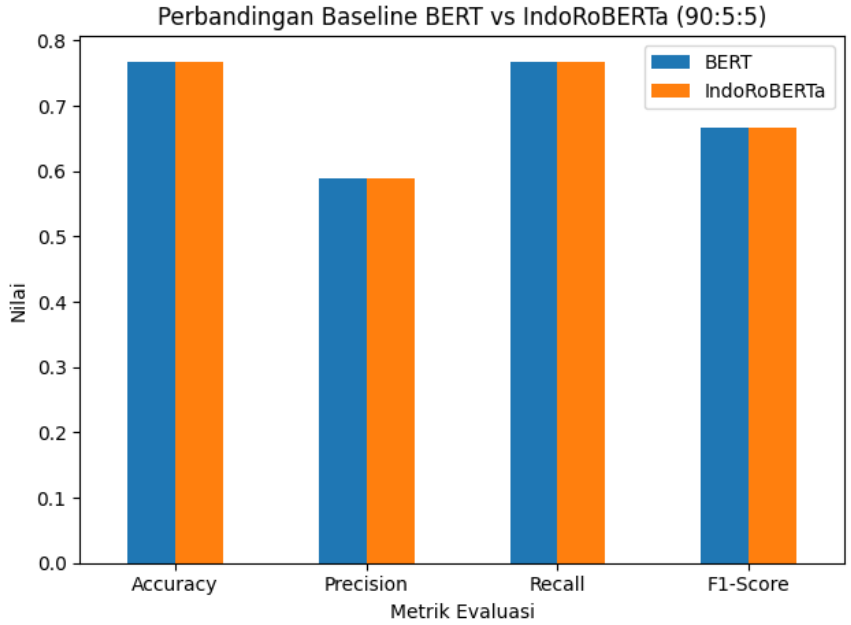
Hasil pengujian awal pada skenario ini disajikan pada Tabel 4.21.

Tabel 4. 21 *Baseline* Skenario 90:5:5

Model	Accuracy	Precision	Recall	F1-Score
BERT	0.77	0.59	0.77	0.67
IndoRoBERTa	0.77	0.59	0.77	0.67

Berdasarkan Tabel 4.21, kedua model kembali menunjukkan performa yang identik pada tahap *baseline* dengan akurasi sebesar 0,77. Nilai *precision* dan *F1-score* yang relatif rendah menunjukkan bahwa model masih didominasi oleh prediksi pada kelas mayoritas (netral), sehingga belum mampu mengklasifikasikan kelas minoritas secara optimal. Hal ini menegaskan bahwa tanpa proses fine-tuning, model belum mampu menangani ketidakseimbangan data.

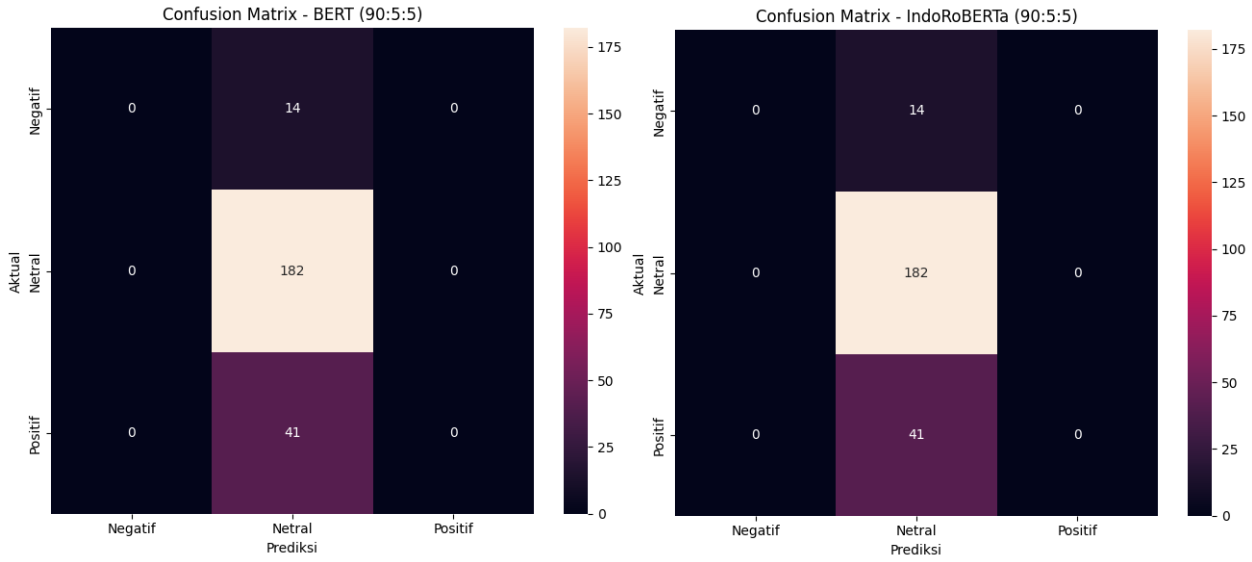
Untuk memberikan gambaran yang lebih jelas mengenai perbandingan performa kedua model, hasil *baseline* juga divisualisasikan dalam bentuk grafik sebagaimana ditunjukkan pada Gambar 4.11.



Gambar 4. 11 Perbandingan *Baseline* BERT dan IndoRoBERTa (90:5:5)

Berdasarkan Gambar 4.11, terlihat bahwa kedua model memiliki nilai evaluasi yang sama pada seluruh metrik. Nilai *accuracy* dan *recall* masing-masing sebesar 0,77, *precision* sebesar 0,59, serta *f1-score* sebesar 0,67. Hasil ini menunjukkan bahwa pada tahap awal, performa BERT dan IndoRoBERTa masih belum optimal dalam melakukan klasifikasi multikelas, terutama pada kelas dengan jumlah data yang lebih sedikit. Oleh karena itu, diperlukan optimasi lanjutan melalui *preprocessing*, penyesuaian *hyperparameter*, dan penanganan ketidakseimbangan data agar performa model dapat meningkat secara lebih merata pada seluruh kelas sentimen.

Untuk melihat distribusi prediksi benar dan salah secara lebih rinci, hasil evaluasi *baseline* BERT dan IndoRoBERTa juga disajikan dalam satu *confusion matrix* gabungan pada Gambar 4.12. Visualisasi ini membantu menunjukkan pola klasifikasi model yang tidak terlihat secara langsung dari nilai *accuracy*, *precision*, *recall*, dan *f1-score*.



Gambar 4. 12 *Confusion Matrix baseline* BER dan IndoRoBERTa (90:5:5)

Berdasarkan Gambar 4.12, terlihat bahwa model BERT dan IndoRoBERTa memiliki pola prediksi yang sama, yaitu seluruh data uji diprediksi ke dalam kelas netral. Sebanyak 182 data netral berhasil diprediksi dengan benar sebagai netral, sedangkan 14 data negatif dan 41 data positif juga seluruhnya diprediksi sebagai netral. Tidak terdapat prediksi pada kelas negatif maupun positif.

Kondisi ini menunjukkan bahwa kedua model hanya mampu mengenali kelas mayoritas dan belum dapat membedakan kelas sentimen lainnya secara optimal. Akibatnya, nilai *precision*, *recall*, dan *f1-score* pada kelas minoritas menjadi sangat rendah. Hal ini mengindikasikan bahwa pada tahap *baseline* masih terjadi *class imbalance*, sehingga model cenderung bias terhadap kelas netral dan memerlukan proses *fine-tuning* untuk meningkatkan performa klasifikasi secara lebih merata.

b. *Fine Tuning*

Pada skenario pembagian data 90:5:5, proses *fine-tuning* dilakukan menggunakan proporsi data latih yang lebih besar dibandingkan skenario sebelumnya, sehingga model memiliki kesempatan untuk mempelajari pola data secara lebih mendalam. Selain itu, penerapan *class weight* tetap digunakan untuk

mengatasi ketidakseimbangan kelas dan membantu model mengenali seluruh kelas sentimen secara lebih seimbang.

Pelatihan dilakukan selama 5 *epoch* dengan evaluasi menggunakan metrik *training loss, validation loss, accuracy, precision, recall,* dan *F1-score*. Hasil *training* model BERT pada setiap *epoch* ditunjukkan pada Tabel 4.22.

Tabel 4. 22 Hasil Training BERT per Epoch Skenario 90:5:5

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1-Score
1	0.908067	0.944346	0.817797	0.791169	0.817797	0.801328
2	0.675980	0.601404	0.851695	0.884885	0.851695	0.862313
3	0.362320	0.534813	0.923729	0.923990	0.923729	0.922922
4	0.237687	0.630839	0.932203	0.932018	0.932203	0.930878
5	0.156120	0.607828	0.940678	0.940875	0.940678	0.938973

Berdasarkan Tabel 4.22, model BERT menunjukkan peningkatan performa yang konsisten selama proses *fine-tuning*. Nilai *training loss* menurun dari 0.908067 pada *epoch* pertama menjadi 0.156120 pada *epoch* kelima, yang menunjukkan bahwa model semakin baik dalam mempelajari pola data latih. Nilai *validation loss* juga menurun dari 0.944346 menjadi 0.607828 meskipun sempat meningkat pada *epoch* keempat.

Akurasi model meningkat dari 0.817797 pada *epoch* pertama menjadi 0.940678 pada *epoch* kelima. Nilai *precision* juga naik dari 0.791169 menjadi 0.940875, *recall* dari 0.817797 menjadi 0.940678, dan *f1-score* dari 0.801328 menjadi 0.938973. Performa terbaik dicapai pada *epoch* kelima, yang menunjukkan bahwa penambahan proporsi data latih pada skenario 90:5:5 memberikan pengaruh positif terhadap kemampuan klasifikasi model BERT.

Sedangkan hasil training model IndoRoBERTa disajikan pada Tabel 4.23.

Tabel 4. 23 Hasil Training IndoRoBERTa per Epoch Skenario 90:5:5

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1-Score
1	0.752605	0.498116	0.927966	0.928314	0.927966	0.926576
2	0.391644	0.188990	0.966102	0.966904	0.966102	0.966416
3	0.155680	0.374510	0.957627	0.957480	0.957627	0.956613
4	0.097224	0.501149	0.961864	0.961169	0.961864	0.960529
5	0.070128	0.442616	0.974576	0.975388	0.974576	0.973712

Berdasarkan Tabel 4.23, model IndoRoBERTa menunjukkan peningkatan performa yang konsisten selama proses *fine-tuning*. Nilai *training loss* menurun dari 0.752605 menjadi 0.070128, sedangkan *accuracy* meningkat dari 0.927966 menjadi 0.974576. Nilai *precision*, *recall*, dan *f1-score* juga meningkat secara signifikan. *Validation loss* terendah terjadi pada *epoch* kedua sebesar 0.188990, namun meningkat pada *epoch* berikutnya yang mengindikasikan *overfitting* ringan. Performa terbaik dicapai pada *epoch* kelima dengan *accuracy* sebesar 0.974576 dan *f1-score* sebesar 0.973712, menunjukkan bahwa IndoRoBERTa memiliki proses pembelajaran yang cepat dan stabil.

c. Hasil Evaluasi Akhir

Hasil evaluasi akhir pada skenario ini disajikan pada Tabel 4.24.

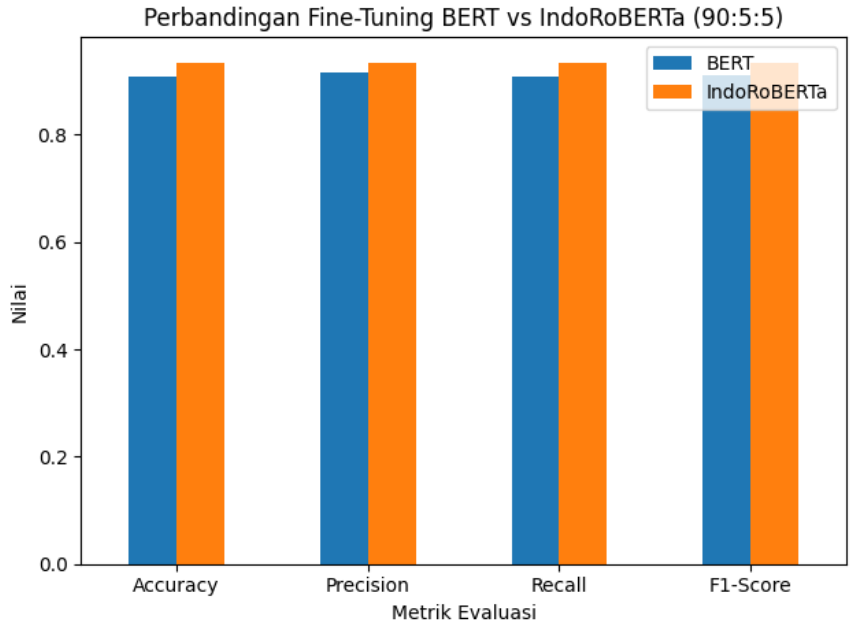
Tabel 4. 24 Hasil Evaluasi Skenario 90:5:5

Model	Accuracy	Precision	Recall	F1-Score
Fine-Tuning BERT	0.91	0.92	0.91	0.91
Fine-Tuning IndoRoBERTa	0.93	0.93	0.93	0.93

Berdasarkan Tabel 4.24, model IndoRoBERTa menunjukkan performa yang lebih baik dibandingkan BERT pada tahap evaluasi akhir skenario 90:5:5. Berdasarkan hasil perhitungan *weighted average*, model BERT memperoleh *accuracy* sebesar 0.907173, *precision* sebesar 0.915006, *recall* sebesar 0.907173, dan *f1-score* sebesar 0.909341 yang dibulatkan menjadi 0.91, 0.92, 0.91, dan 0.91. Sementara itu, model IndoRoBERTa memperoleh *accuracy* sebesar 0.932489, *precision* sebesar 0.934477, *recall* sebesar 0.932489, dan *f1-score* sebesar 0.933193 yang dibulatkan menjadi 0.93 pada seluruh metrik evaluasi. Hasil ini menunjukkan bahwa IndoRoBERTa memiliki performa klasifikasi yang lebih stabil dan lebih baik dibandingkan BERT. Nilai *precision* dan *f1-score* yang lebih tinggi menunjukkan bahwa IndoRoBERTa lebih mampu menjaga keseimbangan antara ketepatan prediksi dan kemampuan mengenali data aktual pada seluruh kelas sentimen.

Meskipun skenario 90:5:5 menggunakan jumlah data latih terbesar, hasil evaluasi tidak menunjukkan peningkatan performa yang lebih tinggi dibandingkan skenario 80:10:10. Hal ini mengindikasikan bahwa penambahan jumlah data latih tidak selalu menghasilkan peningkatan kinerja secara langsung. Ukuran data validasi dan data uji yang lebih kecil dapat menyebabkan hasil evaluasi menjadi lebih sensitif terhadap variasi data sehingga kemampuan generalisasi model tidak tergambar secara optimal.

Untuk memperjelas perbandingan performa akhir setelah proses *fine-tuning*, hasil evaluasi model BERT dan IndoRoBERTa disajikan dalam bentuk grafik pada Gambar 4.13. Grafik tersebut menampilkan nilai *accuracy*, *precision*, *recall*, dan *F1-score* sehingga perbedaan kinerja kedua model dapat diamati dengan lebih jelas.



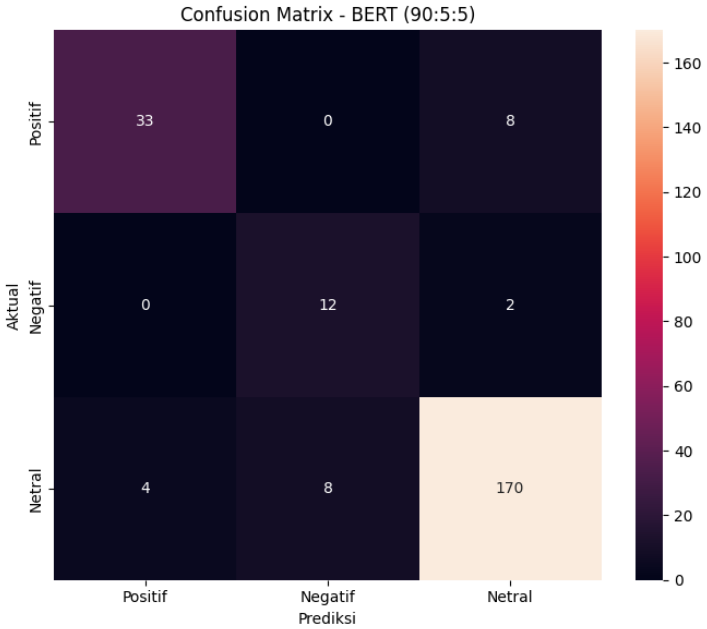
Gambar 4. 13 Perbandingan *Fine-Tuning* BERT dan IndoRoBERTa (90:5:5)

Berdasarkan Gambar 4.13, kedua model menunjukkan performa yang sangat tinggi setelah proses *fine-tuning*, dengan seluruh metrik berada di atas 0,90. Secara rinci, pada metrik *accuracy*, BERT memperoleh nilai 0,90, sedangkan IndoRoBERTa mencapai 0,92. Pada *precision*, BERT berada pada 0,91, sementara IndoRoBERTa sedikit lebih tinggi yaitu 0,92.

Selanjutnya, pada *recall*, BERT mencatat nilai 0,90, sedangkan IndoRoBERTa mencapai 0,92, yang menunjukkan kemampuan lebih baik dalam menangkap seluruh data yang relevan. Pada metrik *F1-score*, BERT memperoleh 0,90, sementara IndoRoBERTa kembali unggul dengan nilai 0,92.

Secara keseluruhan, meskipun selisihnya relatif kecil (sekitar 0,01–0,02), IndoRoBERTa secara konsisten mengungguli BERT pada seluruh metrik evaluasi, yang mengindikasikan performa yang lebih stabil dan lebih optimal dalam klasifikasi sentimen pada skenario pembagian data 90:5:5.

Untuk memberikan gambaran yang lebih rinci mengenai kinerja model, hasil klasifikasi BERT setelah proses *fine-tuning* divisualisasikan dalam bentuk *confusion matrix* pada Gambar 4.14, yang menunjukkan distribusi prediksi pada kelas negatif, netral, dan positif.

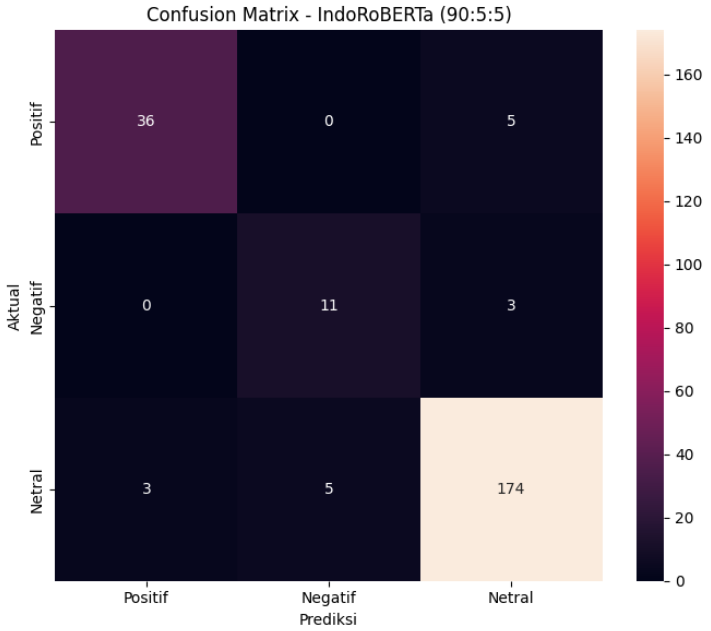


Gambar 4. 14 *Confusion Matrix Fine-Tuning* BERT (90:5:5)

Berdasarkan Gambar 4.14, distribusi prediksi menunjukkan bahwa sebagian besar data berhasil diklasifikasikan dengan benar oleh model, di mana pada kelas positif terdapat 33 data yang diprediksi benar sebagai positif, tidak ada yang salah ke negatif (0), dan 8 data keliru diprediksi sebagai netral; pada kelas negatif, sebanyak 12 data berhasil diklasifikasikan dengan tepat sebagai negatif, tanpa kesalahan ke positif (0), namun terdapat 2 data yang salah diprediksi

sebagai netral; sedangkan pada kelas netral, performa model paling dominan dengan 170 data terklasifikasi benar sebagai netral, meskipun masih terdapat kesalahan prediksi ke positif sebanyak 4 data dan ke negatif sebanyak 8 data, sehingga secara keseluruhan kesalahan paling sering terjadi pada pergeseran ke kelas netral yang mengindikasikan adanya kecenderungan model dalam mengklasifikasikan data yang ambigu sebagai netral

Selanjutnya, hasil evaluasi *fine-tuning* model IndoRoBERTa ditampilkan dalam grafik pada Gambar 4.15 yang memuat *accuracy*, *precision*, *recall*, dan *F1-score* untuk memperjelas kinerjanya.



Gambar 4. 15 *Confusion Matrix Fine-Tuning* IndoRoBERTa (90:5:5)

Berdasarkan Gambar 4.15, model IndoRoBERTa menunjukkan kinerja yang cukup baik. Pada kelas positif, dari 41 data aktual, sebanyak 36 berhasil diprediksi dengan benar, sementara 5 data keliru diprediksi sebagai netral. Pada kelas negatif, dari 14 data, 11 diprediksi benar dan 3 lainnya salah ke netral. Adapun pada kelas netral, dari 182 data, sebanyak 174 diprediksi tepat, sedangkan masing-masing 3 dan 5 data keliru ke positif dan negatif. Hal ini menunjukkan bahwa model paling optimal dalam mengenali kelas netral, sementara kesalahan prediksi umumnya cenderung bergeser ke kelas netral.

4.5.4 Perbandingan Antar Skenario

Untuk memberikan gambaran menyeluruh terhadap kinerja model pada setiap skenario, dilakukan perbandingan hasil evaluasi yang disajikan pada Tabel 4.25.

Tabel 4. 25 Perbandingan Seluruh Skenario

Skenario	Model	Accuracy	Precision	Recall	F1-Score
70:15:15	BERT	0.887165	0.887772	0.887165	0.887272
70:15:15	IndoRoBERTa	0.925247	0.924481	0.925247	0.924733
80:10:10	BERT	0.945032	0.948138	0.945032	0.946072
80:10:10	IndoRoBERTa	0.945032	0.945071	0.945032	0.944768
90:5:5	BERT	0.907173	0.915006	0.907173	0.909341
90:5:5	IndoRoBERTa	0.932489	0.934477	0.932489	0.933193

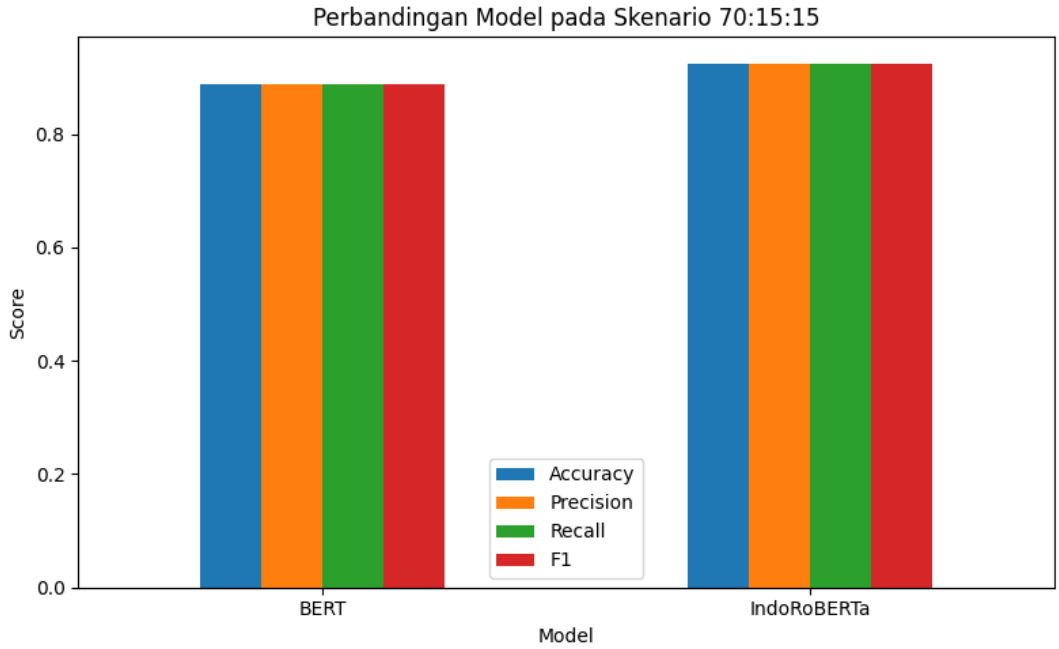
Berdasarkan Tabel 4.25, performa model BERT dan IndoRoBERTa menunjukkan hasil yang berbeda pada setiap skenario pembagian data. Pada skenario 70:15:15, IndoRoBERTa memperoleh hasil yang lebih baik dibandingkan BERT dengan *accuracy* sebesar 0.925247 dan *f1-score* sebesar 0.924733, sedangkan BERT memperoleh *accuracy* sebesar 0.887165 dan *f1-score* sebesar 0.887272. Pada skenario 80:10:10, kedua model memiliki nilai *accuracy* yang sama sebesar 0.945032, namun BERT sedikit lebih unggul pada *precision* dan *f1-score*, yaitu 0.948138 dan 0.946072 dibandingkan IndoRoBERTa yang memperoleh 0.945071 dan 0.944768.

Sementara itu, pada skenario 90:5:5, IndoRoBERTa kembali menunjukkan performa yang lebih baik dengan *accuracy* sebesar 0.932489 dan *f1-score* sebesar 0.933193, sedangkan BERT memperoleh *accuracy* sebesar 0.907173 dan *f1-score* sebesar 0.909341. Secara keseluruhan, IndoRoBERTa menunjukkan performa yang lebih konsisten dan unggul pada sebagian besar skenario, terutama pada pembagian data 70:15:15 dan 90:5:5, sedangkan performa terbaik BERT terdapat pada skenario 80:10:10.

Hasil tersebut menunjukkan bahwa IndoRoBERTa lebih efektif dalam melakukan klasifikasi sentimen komentar publik berbahasa Indonesia karena memiliki kemampuan representasi bahasa yang lebih sesuai dengan karakteristik

data penelitian. Selain memperoleh nilai evaluasi yang lebih tinggi pada sebagian besar skenario, IndoRoBERTa juga menunjukkan kestabilan performa yang lebih baik dibandingkan BERT. Oleh karena itu, IndoRoBERTa dapat dianggap sebagai model yang paling optimal untuk digunakan dalam penelitian ini.

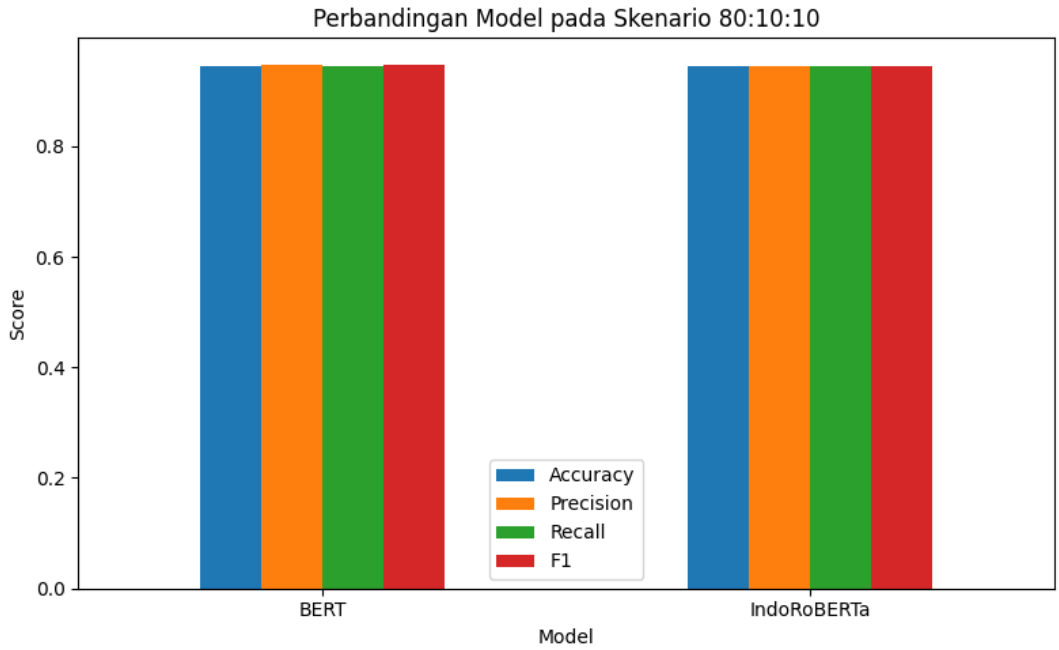
Untuk memperjelas perbandingan performa model pada skenario 70:15:15, hasil evaluasi akhir divisualisasikan dalam bentuk grafik pada Gambar 4.16.



Gambar 4. 16 Perbandingan Model pada Skenario 70:15:15

Berdasarkan gambar 4.16, model BERT memperoleh nilai *accuracy* sebesar 0,89, *precision* sebesar 0,89, *recall* sebesar 0,89, dan *F1-score* sebesar 0,89. Sementara itu, IndoRoBERTa memperoleh nilai *accuracy* sebesar 0,93, *precision* sebesar 0,92, *recall* sebesar 0,93, dan *F1-score* sebesar 0,92. Hasil ini menunjukkan bahwa IndoRoBERTa memiliki performa yang lebih baik dibandingkan BERT pada seluruh metrik evaluasi, sehingga lebih optimal dalam melakukan klasifikasi sentimen pada skenario ini.

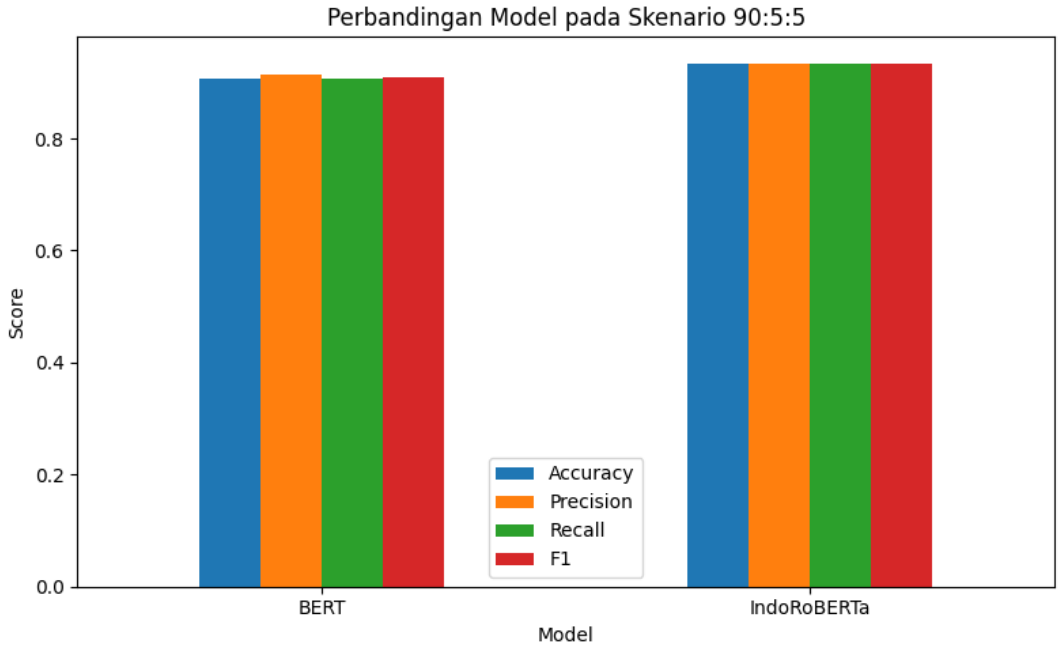
Selanjutnya, pada skenario 80:10:10, hasil evaluasi akhir kedua model disajikan dalam bentuk grafik pada Gambar 4.17 untuk melihat tingkat konsistensi performa setelah proses *fine-tuning*.



Gambar 4. 17 Perbandingan Model pada Skenario 80:10:10

Berdasarkan Gambar 4.17, kedua model menunjukkan performa yang hampir sama. Model BERT memperoleh nilai *accuracy*, *precision*, dan *recall* sebesar 0,95 dengan *F1-score* sebesar 0,95, sedangkan IndoRoBERTa memperoleh nilai *accuracy*, *precision*, dan *recall* sebesar 0,95 dengan *F1-score* sebesar 0,94. Perbedaan performa pada skenario ini sangat kecil, meskipun BERT sedikit lebih unggul pada nilai *precision* dan *F1-score*.

Pada skenario terakhir, yaitu 90:5:5, perbandingan performa model ditampilkan pada Gambar 4.18 untuk melihat pengaruh proporsi data latih yang lebih besar terhadap hasil klasifikasi.



Gambar 4. 18 Perbandingan Model pada Skenario 90:5:5

Berdasarkan Gambar 4.18, model BERT memperoleh nilai *accuracy* sebesar 0,91, *precision* sebesar 0,92, *recall* sebesar 0,91, dan *F1-score* sebesar 0,91. Sementara itu, IndoRoBERTa memperoleh nilai *accuracy*, *precision*, *recall*, dan *F1-score* sebesar 0,93. Hasil ini menunjukkan bahwa IndoRoBERTa kembali memiliki performa yang lebih baik dibandingkan BERT, terutama dalam menjaga kestabilan klasifikasi pada seluruh kelas sentimen, sehingga menjadi model yang paling optimal pada skenario ini.

4.5.5 Hasil Uji Statistik Paired Sample t-Test

Setelah dilakukan perbandingan performa model BERT dan IndoRoBERTa pada seluruh skenario pembagian data, dilakukan pengujian statistik menggunakan *paired sample t-test* untuk mengetahui apakah perbedaan performa kedua model bersifat signifikan secara statistik. Pengujian dilakukan terhadap nilai *accuracy* dan *F1-score* yang diperoleh pada tiga skenario penelitian, yaitu 70:15:15, 80:10:10, dan 90:5:5.

Hipotesis yang digunakan dalam penelitian ini adalah sebagai berikut:

- H_0: Tidak terdapat perbedaan signifikan antara performa BERT dan IndoRoBERTa.

- H_1: Terdapat perbedaan signifikan antara performa BERT dan IndoRoBERTa.

Tingkat signifikansi yang digunakan adalah $\alpha = 0{,}05$. Apabila nilai *p-value* < 0,05 maka H_0 ditolak dan H_1 diterima. Sebaliknya, apabila *p-value* ≥ 0,05 maka H_0 gagal ditolak.

Hasil pengujian *paired sample t-test* disajikan pada Tabel 4.26.

Tabel 4. 26 Hasil Uji *Paired Sample t-Test*

Metrik	t-statistic	p-value	Keputusan	Metrik	t-statistic
Accuracy	-1.888	0.200	H_0 diterima	Accuracy	-1.888
F1-Score	-1.762	0.220	H_0 diterima	F1-Score	-1.762

Berdasarkan Tabel 4.26, nilai *p-value* untuk metrik *accuracy* sebesar 0,200 dan *F1-score* sebesar 0,220. Kedua nilai tersebut lebih besar daripada tingkat signifikansi 0,05, sehingga H_0 gagal ditolak. Hasil ini menunjukkan bahwa perbedaan performa antara model BERT dan IndoRoBERTa tidak signifikan secara statistik berdasarkan nilai *accuracy* maupun *F1-score*.

Meskipun pada beberapa skenario IndoRoBERTa memperoleh nilai evaluasi yang lebih tinggi dibandingkan BERT, terutama pada skenario 70:15:15 dan 90:5:5, selisih performa yang diperoleh belum cukup besar untuk menunjukkan adanya perbedaan yang signifikan secara statistik. Dengan demikian, kedua model dapat dianggap memiliki performa yang relatif setara dalam melakukan klasifikasi sentimen komentar YouTube pada dataset penelitian ini.

Hasil pengujian statistik ini memperkuat temuan pada tahap evaluasi sebelumnya, di mana IndoRoBERTa cenderung menunjukkan nilai *accuracy* dan *F1-score* yang lebih tinggi pada sebagian besar skenario. Namun, berdasarkan uji signifikansi statistik, keunggulan tersebut belum dapat dinyatakan berbeda secara signifikan pada tingkat kepercayaan 95%. Oleh karena itu, pemilihan model terbaik dalam penelitian ini lebih didasarkan pada konsistensi nilai evaluasi yang diperoleh pada setiap skenario, di mana IndoRoBERTa tetap menunjukkan performa yang lebih stabil dibandingkan BERT.

Hasil ini menunjukkan bahwa peningkatan nilai evaluasi tidak selalu diikuti oleh perbedaan yang signifikan secara statistik. Dalam konteks klasifikasi sentimen, kondisi tersebut mengindikasikan bahwa kedua model sama-sama memiliki kemampuan yang baik dalam mengenali pola sentimen pada komentar YouTube. Meskipun demikian, IndoRoBERTa menunjukkan kecenderungan performa yang lebih konsisten pada sebagian besar skenario pengujian, sehingga model ini dinilai lebih sesuai untuk digunakan pada klasifikasi sentimen berbahasa Indonesia dalam penelitian ini.

4.6 Relevansi Penelitian dengan Perspektif Islam

Berdasarkan temuan penelitian, analisis sentimen terhadap komentar publik di YouTube mengenai kecerdasan buatan (AI) dalam pendidikan memperlihatkan bahwa mayoritas komentar berada pada kategori netral, kemudian diikuti oleh sentimen positif dan negatif. Dominasi sentimen netral menunjukkan bahwa masyarakat cenderung memandang AI secara rasional, yaitu sebagai teknologi yang memiliki manfaat besar, namun tetap memunculkan kehati-hatian terhadap risiko yang mungkin ditimbulkan. Hal ini menandakan bahwa penerimaan masyarakat terhadap AI tidak bersifat mutlak, melainkan dipengaruhi oleh pertimbangan manfaat, etika, dan dampaknya dalam kehidupan sosial maupun pendidikan.

Selain itu, hasil evaluasi model menunjukkan bahwa IndoRoBERTa memperoleh performa yang lebih baik dibandingkan BERT pada sebagian besar skenario pembagian data. Pada skenario 70:15:15, IndoRoBERTa memperoleh *accuracy* sebesar 0,93 dan *F1-score* sebesar 0,92; pada skenario 80:10:10 memperoleh *accuracy* sebesar 0,95 dan *F1-score* sebesar 0,94; serta pada skenario 90:5:5 memperoleh *accuracy* sebesar 0,93 dan *F1-score* sebesar 0,93. Hasil tersebut menunjukkan bahwa IndoRoBERTa lebih optimal dalam memahami sentimen masyarakat terhadap AI karena memiliki kemampuan representasi bahasa yang lebih sesuai dengan karakteristik bahasa Indonesia. Temuan ini juga memperlihatkan bahwa teknologi dapat dimanfaatkan untuk membantu

pengambilan keputusan yang lebih akurat, khususnya dalam bidang pendidikan dan analisis kebijakan publik.

Dalam pandangan Islam, perkembangan teknologi seperti AI merupakan bagian dari pemanfaatan akal dan ilmu yang Allah SWT anugerahkan kepada manusia. Hal tersebut ditegaskan dalam QS. Al-Baqarah [2]: 31:

وَعَلَّمَ اٰدَمَ الْاَسْمَاۤءَ كُلَّهَا ثُمَّ عَرَضَهُمْ عَلَى الْمَلٰۤىِٕكَةِ فَقَالَ اَنْۢبِـُٔوْنِيْ بِاَسْمَاۤءِ هٰٓؤُلَاۤءِ اِنْ كُنْتُمْ صٰدِقِيْنَ ﴿٣١﴾

Artinya: *"Dia mengajarkan kepada Adam nama-nama (benda) seluruhnya, kemudian Dia memperlihatkannya kepada para malaikat, seraya berfirman, 'Sebutkan kepada-Ku nama-nama (benda) ini jika kamu benar!'"* (QS. Al-Baqarah [2]: 31); terjemahan Qur'an NU Online, 2026).

Menurut Tafsir al-Munir, ayat tersebut menjelaskan bahwa manusia memiliki keistimewaan dalam penguasaan ilmu pengetahuan yang menjadi landasan utama dalam pengembangan sains dan teknologi, termasuk AI yang merepresentasikan proses berpikir logis dan pengambilan keputusan berbasis data (Alfiansyah *et al*., 2023). Dalam penelitian ini, penggunaan model BERT dan IndoRoBERTa untuk klasifikasi sentimen menunjukkan bagaimana ilmu pengetahuan modern dimanfaatkan untuk memahami respons masyarakat secara sistematis dan terukur. Dengan demikian, teknologi AI dapat dipahami sebagai instrumen yang membantu manusia dalam menghasilkan keputusan yang lebih objektif.

Prinsip tersebut juga diperkuat oleh QS. Al-'Alaq [1–5]:

اقْرَأْ بِاسْمِ رَبِّكَ الَّذِي خَلَقَ (١) خَلَقَ الْإِنسَانَ مِنْ عَلَقٍ (٢) اقْرَأْ وَرَبُّكَ الْأَكْرَمُ (٣) الَّذِي عَلَّمَ بِالْقَلَمِ (٤) عَلَّمَ الْإِنسَانَ مَا لَمْ يَعْلَمْ (٥)

Artinya: *"Bacalah dengan (menyebut) nama Tuhanmu yang menciptakan!. Dia menciptakan manusia dari segumpal darah. Bacalah! Tuhanmulah Yang Mahamulia. Yang mengajar (manusia) dengan pena. Dia mengajarkan manusia apa yang tidak diketahuinya."* (QS. Al-'Alaq [96]: 1–5; terjemahan Qur'an NU Online, 2026).

Ayat ini menegaskan bahwa pengembangan teknologi harus dilandasi oleh ilmu pengetahuan, etika, dan orientasi kemaslahatan, sejalan dengan konsep *maqāṣid al-sharīʿah* yang menempatkan perlindungan akal dan kesejahteraan manusia sebagai dasar moral dalam penggunaan AI (Masykur & Solekhah, 2021).

Berdasarkan hasil penelitian, banyak komentar positif yang menunjukkan bahwa AI dipandang mampu membantu mahasiswa dan pendidik dalam proses belajar, pencarian referensi, serta peningkatan efisiensi akademik. Kondisi ini memperlihatkan bahwa AI dapat menjadi sarana kemanfaatan apabila digunakan secara tepat dan bertanggung jawab.

Di sisi lain, penelitian ini juga menemukan adanya sentimen negatif yang berkaitan dengan kekhawatiran masyarakat terhadap ketergantungan pada AI, kemungkinan penyalahgunaan informasi, dan ancaman berkurangnya peran manusia dalam proses pendidikan. Fenomena ini relevan dengan QS. Al-Isra' [17]: 36:

وَلَا تَقْفُ مَا لَيْسَ لَكَ بِهٖ عِلْمٌۗ اِنَّ السَّمْعَ وَالْبَصَرَ وَالْفُؤَادَ كُلُّ اُولٰۤىِٕكَ كَانَ عَنْهُ مَسْـُٔوْلًا ﴿٣٦﴾

Artinya: *"Janganlah engkau mengikuti sesuatu yang tidak kauketahui. Sesungguhnya pendengaran, penglihatan, dan hati nurani, semua itu akan diminta pertanggungjawabannya."* (QS. Al-Isra' [17]: 36; terjemahan Qur'an NU Online, 2026).

Ayat ini menekankan pentingnya verifikasi informasi dan tanggung jawab moral dalam menerima maupun menggunakan pengetahuan, yang sangat relevan dengan etika pemanfaatan AI (Ali et al., 2025). Dalam konteks penelitian ini, hasil analisis menunjukkan bahwa pengguna AI tidak seharusnya menerima seluruh informasi secara langsung tanpa proses validasi, terutama dalam bidang pendidikan yang menuntut ketepatan dan akurasi ilmiah.

Prinsip kehati-hatian tersebut juga diperkuat dalam QS. Al-Hujurat [49]: 6:

يٰٓاَيُّهَا الَّذِيْنَ اٰمَنُوْٓا اِنْ جَاۤءَكُمْ فَاسِقٌۢ بِنَبَاٍ فَتَبَيَّنُوْٓا اَنْ تُصِيْبُوْا قَوْمًاۢ بِجَهَالَةٍ فَتُصْبِحُوْا عَلٰى مَا فَعَلْتُمْ نٰدِمِيْنَ ﴿٦﴾

Artinya: *"Wahai orang-orang yang beriman, jika seorang fasik datang kepadamu membawa berita penting, maka telitilah kebenarannya agar kamu tidak mencelakakan suatu kaum karena ketidaktahuan(-mu) yang berakibat kamu menyesali perbuatanmu itu."* (QS. Al-Hujurat [49]: 6; terjemahan Qur'an NU Online, 2026).

Ayat tersebut sangat relevan dengan fenomena digital saat ini, khususnya terkait validitas informasi yang dihasilkan AI serta potensi penyebaran misinformasi (Molla & Ahsan, 2025). Temuan penelitian menunjukkan bahwa sebagian masyarakat masih meragukan ketepatan jawaban AI, sehingga teknologi

ini perlu diposisikan sebagai alat bantu, bukan sebagai sumber kebenaran utama yang diterima tanpa kritik.

Selain itu, dalam hadis riwayat Muslim disebutkan:

عَنْ أَبِي هُرَيْرَةَ أَنَّ رَسُولَ اللهِ ﷺ قَالَ مَنْ دَعَا إِلَى هُدًى كَانَ لَهُ مِنَ الأَجْرِ مِثْلُ أُجُورِ مَنْ تَبِعَهُ لاَ يَنْقُصُ ذَلِكَ مِنْ أُجُورِهِمْ شَيْئًا وَمَنْ دَعَا إِلَى ضَلاَلَةٍ كَانَ عَلَيْهِ مِنَ الإِثْمِ مِثْلُ آثَامِ مَنْ تَبِعَهُ لاَ يَنْقُصُ ذَلِكَ مِنْ آثَامِهِمْ شَيْئًا

Artinya: *"Barang siapa yang menunjukkan kepada kebaikan, maka ia mendapatkan pahala seperti orang yang melakukannya."* (HR. Muslim).

Hadis ini memberikan penegasan bahwa penggunaan teknologi dapat bernilai ibadah apabila diarahkan untuk membawa manfaat bagi masyarakat. Dalam penelitian ini, model IndoRoBERTa menunjukkan performa yang lebih baik dibandingkan BERT dalam memahami sentimen publik terhadap AI. Hasil tersebut menunjukkan bahwa teknologi dapat dimanfaatkan untuk membantu pengambilan keputusan yang lebih akurat, khususnya dalam bidang pendidikan dan perumusan kebijakan publik.

Dalam etika Islam, pemanfaatan AI harus mengutamakan kemaslahatan dan mencegah kerusakan, sebagaimana kaidah fiqh *Dar'u al-mafāsid muqaddam 'alā jalb al-maṣāliḥ* serta prinsip *maqāṣid al-sharī'ah* (Ali et al., 2025). Temuan penelitian ini menunjukkan bahwa masyarakat menerima AI selama teknologi tersebut digunakan untuk mendukung kualitas pendidikan, bukan untuk menggantikan sepenuhnya peran manusia.

Hal ini sejalan dengan konsep manusia sebagai khalifah di bumi sebagaimana dijelaskan dalam QS. Al-Baqarah [2]: 30:

وَاِذْ قَالَ رَبُّكَ لِلْمَلٰۤىِٕكَةِ اِنِّيْ جَاعِلٌ فِى الْاَرْضِ خَلِيْفَةً ۗقَالُوْٓا اَتَجْعَلُ فِيْهَا مَنْ يُّفْسِدُ فِيْهَا وَيَسْفِكُ الدِّمَاۤءَ ۚوَنَحْنُ نُسَبِّحُ بِحَمْدِكَ وَنُقَدِّسُ لَكَ ۗقَالَ اِنِّيْٓ اَعْلَمُ مَا لَا تَعْلَمُوْنَ ۝٣٠

Artinya: *"(Ingatlah) ketika Tuhanmu berfirman kepada para malaikat, "Aku hendak menjadikan khalifah di bumi." Mereka berkata, "Apakah Engkau hendak menjadikan orang yang merusak dan menumpahkan darah di sana, sedangkan kami bertasbih memuji-Mu dan menyucikan nama-Mu?" Dia berfirman, "Sesungguhnya Aku mengetahui apa yang tidak kamu ketahui."* (QS. Al-Baqarah [2]: 30; terjemahan Qur'an NU Online, 2026).

Ayat ini menjelaskan bahwa manusia memiliki amanah moral dalam mengelola ilmu pengetahuan dan teknologi agar tidak menimbulkan kerusakan

sosial. Oleh karena itu, hasil penelitian ini menegaskan bahwa AI dalam pendidikan dapat diterima dalam perspektif Islam selama penggunaannya tetap berada dalam koridor ilmu, etika, tanggung jawab, dan kemaslahatan umat.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian mengenai klasifikasi sentimen komentar publik berbahasa Indonesia di YouTube terkait kecerdasan buatan (*AI*) dalam bidang pendidikan, dapat disimpulkan bahwa model BERT dan IndoRoBERTa mampu melakukan klasifikasi sentimen ke dalam kategori positif, netral, dan negatif dengan performa yang baik. Pada tahap *baseline*, kedua model menunjukkan kinerja yang relatif sebanding, sedangkan setelah dilakukan *fine-tuning* terjadi peningkatan performa pada seluruh skenario pembagian data yang digunakan. Hasil pengujian menunjukkan bahwa IndoRoBERTa cenderung menghasilkan nilai *accuracy*, *precision*, *recall*, dan *F1-score* yang lebih tinggi dibandingkan BERT pada sebagian besar skenario, yang mengindikasikan kemampuan lebih baik dalam memahami konteks bahasa Indonesia karena dilatih menggunakan korpus berbahasa Indonesia dalam jumlah besar. Namun demikian, berdasarkan hasil uji statistik *paired sample t-test*, perbedaan performa antara kedua model tidak signifikan secara statistik pada tingkat signifikansi 0,05, sehingga keduanya dapat dianggap memiliki kemampuan yang relatif setara dalam konteks dataset penelitian ini. Secara keseluruhan, pendekatan berbasis *Transformer* terbukti efektif untuk klasifikasi sentimen, dan hasil penelitian menunjukkan bahwa proses *fine-tuning* mampu meningkatkan kualitas klasifikasi dibandingkan model *pre-trained* tanpa pelatihan ulang, sehingga temuan ini dapat menjadi referensi bagi pengembangan penelitian analisis sentimen berbahasa Indonesia serta mendukung pemanfaatan *Natural Language Processing* dalam memahami opini publik secara lebih akurat.

5.2 Saran

Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa saran yang dapat dipertimbangkan untuk pengembangan penelitian selanjutnya, yaitu penggunaan dataset dengan jumlah yang lebih besar serta distribusi kelas yang

lebih seimbang guna meningkatkan kemampuan generalisasi model dan menghasilkan evaluasi yang lebih representatif. Selain itu, sumber data dapat diperluas tidak hanya dari komentar *YouTube*, tetapi juga dari berbagai platform media sosial lainnya agar diperoleh variasi opini yang lebih beragam. Penelitian berikutnya juga disarankan untuk mengeksplorasi model berbasis *Transformer* lain, seperti *IndoBERTweet*, *XLM-RoBERTa*, atau model bahasa Indonesia terbaru sehingga dapat diperoleh perbandingan performa yang lebih komprehensif, disertai pengembangan proses *fine-tuning* melalui optimasi *hyperparameter* yang lebih sistematis untuk memperoleh konfigurasi model yang optimal. Selanjutnya, evaluasi model dapat diperluas dengan menambahkan skenario pengujian yang lebih beragam atau menggunakan teknik validasi seperti *k-fold cross validation* agar hasil yang diperoleh lebih *robust*, serta dilengkapi dengan analisis kesalahan (*error analysis*) pada setiap kelas sentimen untuk mengidentifikasi pola kesalahan klasifikasi dan faktor yang memengaruhinya. Terakhir, penelitian ini dapat dikembangkan menuju implementasi sistem analisis sentimen otomatis untuk memantau opini publik terkait pemanfaatan kecerdasan buatan dalam pendidikan maupun isu lainnya sehingga dapat memberikan manfaat praktis sebagai pendukung pengambilan keputusan berbasis data di berbagai bidang.

DAFTAR PUSTAKA

Abidin, D. Z., Afuan, L., Toscany, A. N., & Nurhadi, N. (2025). A Comprehensive Benchmarking Pipeline for Transformer-Based Sentiment Analysis using Cross-Validated Metrics. *Jurnal Teknik Informatika (Jutif)*, *6*(4), 1797–1810. https://doi.org/10.52436/1.jutif.2025.6.4.4894

Abidin, M. I., & Pamungkas, E. W. (2025). Analisis Sentimen Terhadap Timnas Indonesia Di Piala Asia 2023 Dengan Model Transformer Berbahasa Indonesia. *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab*, *10*(2), 482–496. https://doi.org/10.36341/rabit.v10i2.6142

Achsan, H. T. Y., Suhartanto, H., Wibowo, W. C., Dewi, D. A., & Ismed, K. (2023). Automatic Extraction of Indonesian Stopwords. *International Journal of Advanced Computer Science and Applications*, *14*(2), 166–171. https://scholar.ui.ac.id/en/publications/automatic-extraction-of-indonesian-stopwords/

Adablanu, S., Barman, U., & Das, D. (2025). A novel hybrid adaptive transformer framework with multihead self attention for stroke detection. *Adablanu et al. Discover Neuroscience*. https://doi.org/https://doi.org/10.1186/s13064-025-00224-7

Al-Hasyimi, A. (1996). *Jawahir al-Adab fi Adabiyyat wa Insya' Lughat al-'Arab*. Dar al-Kutub al-Ilmiyyah. https://digilib.walisongo.ac.id/slims/index.php?id=4226&p=show_detail

Alaei, A., Wang, Y., Bui, V., & Stantic, B. (2023). Target-Oriented Data Annotation for Emotion and Sentiment Analysis in Tourism Related Social Media Data. *Future Internet*. https://doi.org/https://doi.org/10.3390/fi15040150

Alawi, A. B., & Bozkurt, F. (2024). A hybrid machine learning model for sentiment analysis and satisfaction assessment with Turkish universities using Twitter data. *Decision Analytics Journal*, *11*(April), 100473. https://doi.org/10.1016/j.dajour.2024.100473

Alfiansyah, M., Khairani, L., Iraya, P., & Hamdani, H. (2023). Istilah Pendidikan

Islam (Ta'lim) Dalam Qs. Al-Baqarah : 31 Menurut Tafsir Al-Munir. *Innovative: Journal Of Social Science Research*, *3*(4), 6105–6116. https://j-innovative.org/index.php/Innovative/article/view/1916

Ali, F., Bouzoubaa, K., Gelli, F., & Hamzi, B. (2025). Islamic Ethics and AI : An Evaluation of Existing Approaches to AI using Trusteeship Ethics. In *Philosophy & Technology* (Vol. 38, Nomor 3). Springer Netherlands. https://doi.org/10.1007/s13347-025-00922-4

Atak, A. (2023). Exploring the sentiment in Borsa Istanbul with deep learning. *Borsa Istanbul Review*, *23*(S2), S84–S95. https://doi.org/10.1016/j.bir.2023.12.010

Atteveldt, W. Van, Velden, M. A. C. G. Van Der, & Boukes, M. (2021). The Validity of Sentiment Analysis : Comparing Manual Annotation , Crowd-Coding , Dictionary Approaches , and Machine Learning Algorithms The Validity of Sentiment Analysis : Comparing Manual Annotation ,. *Communication Methods and Measures*, *15*(2), 121–140. https://doi.org/10.1080/19312458.2020.1869198

Azzahra, F. S. P., & Printo, C. (2025). Systematic Literature Review: Analysis of AI Implementation for Document Verification. *Jurnal Ilmiah Sistem Informasi (JUSI)*, *4*(3), 417–430. https://doi.org/https://doi.org/10.51903/kjjwk708

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

Bal, M., Aydemir, A. G. ı K., & Coşkun, M. (2024). Exploring YouTube content creators ' perspectives on generative AI in language learning : Insights through opinion mining and sentiment analysis. *Plos One*, 1–30. https://doi.org/10.1371/journal.pone.0308096

Bello, A., Ng, S. C., & Leung, M. F. (2023). A BERT Framework to Sentiment Analysis of Tweets. *Sensors*, *23*(1). https://doi.org/10.3390/s23010506

Bettayeb, M., Halawani, Y., Khan, M. U., & Saleh, H. (2024). Efficient memristor accelerator for transformer self-attention functionality. *Scientific Reports*, 1–

15. https://doi.org/https://www.nature.com/articles/s41598-024-75021-z

Bond, M., Khosravi, H., Laat, M. De, Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education : a call for increased ethics , collaboration , and rigour. *International Journal of Educational Technology in Higher Education*. https://doi.org/10.1186/s41239-023-00436-z

Carvalho, M., Pinho, A. J., & Brás, S. (2025). Resampling approaches to handle class imbalance : a review from a data perspective. *Journal of Big Data*. https://doi.org/10.1186/s40537-025-01119-4

Chen, Y., Pu, H., & Qu, Y. (2024). An analysis of attention mechanisms and its variance in transformer. *Proceedings of the 4th International Conference on Signal Processing and Machine Learning*, *0*(2), 164–176. https://doi.org/10.54254/2755-2721/47/20241291

Chew, H. S. J., & Achananuparp, P. (2022). *Perceptions and Needs of Artificial Intelligence in Health Care to Increase Adoption*. *24*, 1–19. https://doi.org/10.2196/32939

Chi, D., Huang, T., Jia, Z., & Zhang, S. (2025). Research on sentiment analysis of hotel review text based on BERT-TCN-BiLSTM-attention model. *Array*, *25*(July 2024), 100378. https://doi.org/10.1016/j.array.2025.100378

Choo, S., & Kim, W. (2023). A study on the evaluation of tokenizer performance in natural language processing. *Applied Artificial Intelligence*, *37*(1). https://doi.org/10.1080/08839514.2023.2175112

Chu, X., Tian, Z., Zhang, B., Wang, X., & Shen, C. (2023). CONDITIONAL POSITIONAL ENCODINGS. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1–19. https://doi.org/https://arxiv.org/abs/2102.10882

Dhendra, & Utomo, V. G. (2025). Benchmarking IndoBERT and Transformer Models for Sentiment Classification on Indonesian E-Government Service Reviews. *Jurnal Transformatika*, *23*(1), 86–95. https://doi.org/10.26623/transformatika.v23i1.12095

El Majid, A. U., & Nuari, R. (2025). Performance Comparison Of BERT Metrics

and Classical Machine Learning Models (SVM,Naive Bayes) for Sentiment Analysis. *INOVTEK Polbeng - Seri Informatika*, *10*(2), 741–752. https://doi.org/10.35314/wmh3rg23

Fachrie, M., Musdholifah, A., & Pulungan, R. (2025). Effectiveness of data resampling and ensemble learning in multiclass imbalance learning. *Artificial Intelligence Review (2025)*. https://doi.org/https://doi.org/10.1007/s10462-025-11357-w

Fitroh, & Hudaya, F. (2023). Systematic Literature Review : Analisis Sentimen Berbasis Deep Learning. *Jurnal Nasional Teknologi dan Sistem Informasi*, *02*, 132–140. https://doi.org/https://doi.org/10.1109/ACCESS.2022.3210182

Frenda, S., Abercrombie, G., Basile, V., Pedrani, A., & Bernardi, D. (2025). Perspectivist approaches to natural language processing : a survey. *Language Resources and Evaluation*, *59*(2), 1719–1746. https://doi.org/10.1007/s10579-024-09766-4

Gardazi, N. M., Daud, A., Malik, M. K., Bukhari, A., Alsahfi, T., & Alshemaimri, B. (2025). BERT applications in natural language processing : a review. *Artificial Intelligence Review*. https://doi.org/https://doi.org/10.1007/s10462-025-11162-5

Geni, L., Yulianti, E., & Sensuse, D. I. (2023). Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using Bert Language Models. *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, *9*(3), 746–757. https://doi.org/10.26555/jiteki.v9i3.26490

Ghosh, N., Santoni, D., Saha, I., & Felici, G. (2025). A review on the applications of Transformer-based language models for nucleotide sequence analysis. *Computational and Structural Biotechnology Journal*, *27*(March), 1244–1254. https://doi.org/10.1016/j.csbj.2025.03.024

Gutierrez-Vasques, X., Bentz, C., Sozinova, O., & Samardžić, T. (2021). From characters to words: The turning point of BPE merges. *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, 3454–3468. https://doi.org/10.18653/v1/2021.eacl-main.302

He, R., Ravula, A., Kanagal, B., & Ainslie, J. (2021). RealFormer : Transformer Likes Residual Attention. *Computational Linguistics*, 929–943. https://doi.org/https://doi.org/10.18653/v1/2021.findings-acl.81

Hessel, J., Holtzman, A., Forbes, M., Le, R., & Yejin, B. (2021). A Reference-free Evaluation Metric for Image Captioning. *Association for Computational Linguistics*, *2014*. https://doi.org/10.18653/v1/2021.emnlp-main.595

Higuchi, Y., Yan, B., Arora, S., Ogawa, T., Kobayashi, T., & Watanabe, S. (2022). BERT Meets CTC: New Formulation of End-to-End Speech Recognition with Pre-trained Masked Language Model. *Findings of the Association for Computational Linguistics: EMNLP 2022*, 5515–5532. https://doi.org/10.18653/v1/2022.findings-emnlp.54

Hirata, E., & Matsuda, T. (2023). Examining logistics developments in post-pandemic Japan through sentiment analysis of Twitter data. *Asian Transport Studies*, *9*(April 2023), 100110. https://doi.org/10.1016/j.eastsj.2023.100110

Hou, J., Katinskaia, A., Vu, A. D., & Yangarber, R. (2023). Effects of sub-word segmentation on performance of transformer language models. *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 7413–7425. https://doi.org/10.18653/v1/2023.emnlp-main.459

Huo, T., Glueck, D. H., Shenkman, E. A., & Muller, K. E. (2023). Stratified split sampling of electronic health records. *BMC Medical Research Methodology*, *0*, 1–7. https://doi.org/https://doi.org/10.1186/s12874-023-01938-0

Islam, S., Xiangdong, L., & Ahmed, J. (2026). BERT : advancements in language understanding for different NLP tasks : challenges and future perspectives. *Journal of Electrical Systems and Information Technology*, *1*, 1–16. https://doi.org/https://doi.org/10.1186/s43067-026-00341-1

Jamil, M., Hadiyanto, H., & Sanjaya, R. (2024). Sentiment Analysis: Classifying Public Comments on YouTube in Disaster Management Simulation in Indonesia Using Naïve Bayes and Support Vector Machine. *Ingénierie des systèmes d information*, *29*(2), 437–446. https://doi.org/10.18280/isi.290205

Jim, J. R., Talukder, M. A. R., Malakar, P., & Kabir, M. M. (2024). Recent advancements and challenges of NLP-based sentiment analysis : A state-of-

the-art review. *Natural Language Processing Journal*, *6*(January), 100059. https://doi.org/10.1016/j.nlp.2024.100059

Joseph, V. R., Vakayil, A., Joseph, V. R., & Vakayil, A. (2022). SPlit : An Optimal Method for Data Splitting SPlit : An Optimal Method for Data Splitting ABSTRACT. *Technometrics*, *0*(0), 1–23. https://doi.org/10.1080/00401706.2021.1921037

Jullfikar, R., Megasari, R., & Sukamto, R. A. (2025). Pengembangan Chatbot Menggunakan BERT untuk Pengelolaan Komunikasi Beasiswa Ikatan Alumni Universitas Pendidikan Indonesia. *Digital Transformation Technology*, *5*(1), 391–400. https://doi.org/10.47709/digitech.v5i1.6367

Khan, J., & Lee, S. (2021). applied sciences Enhancement of Text Analysis Using Context-Aware Normalization of Social Media Informal Text. *Applied Sciences*. https://doi.org/https://doi.org/10.3390/app11178172

Lee, Y., Son, J., & Song, M. (2022). BertSRC : transformer - based semantic relation classification. *BMC Medical Informatics and Decision Making*, *5*, 1–18. https://doi.org/10.1186/s12911-022-01977-5

Li, Z., Zhao, Y., Zhang, X., Han, H., & Huang, C. (2025). Word embedding factor based multi-head attention. *Artificial Intelligence Review*, *January*. https://doi.org/https://doi.org/10.1007/s10462-025-11115-y

Lindén, K., Jauhiainen, T., & Hardwick, S. (2023). FinnSentiment : a Finnish social media corpus for sentiment polarity annotation. *Language Resources and Evaluation*, *57*(2), 581–609. https://doi.org/10.1007/s10579-023-09644-5

Luca, P. De, Chiodo, A., Macario, A., & Siciliano, C. (2021). Semi-Continuous Adsorption Processes with Multi-Walled Carbon Nanotubes for the Treatment of Water Contaminated by an Organic Textile Dye. *Applied Sciences*. https://doi.org/https://doi.org/10.3390/app11041687

Martínez-miwa, C. A., & Castelán, M. (2023). On reliability of annotations in contextual emotion imagery. *Scientific Data*, 1–12. https://doi.org/10.1038/s41597-023-02435-1

Masykur, & Solekhah, S. (2021). TAFSIR QUR'AN SURAH AL-'ALAQ AYAT

1 SAMPAI 5 (Perspektif Ilmu Pendidikan). *Jurnal Studi Keislaman*, *2*(2), 72–87. https://e-journal.stishid.ac.id/index.php/wasathiyah/article/view/123

Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent Neural Networks : A Comprehensive Review of Architectures , Variants , and Applications. *Information*, 1–34. https://doi.org/https://doi.org/10.3390/info15090517

Moghe, N., Steedman, M., & Birch, A. (2021). Cross-lingual Intermediate Fine-tuning improves Dialogue State Tracking. *Proceedings of EMNLP*, 1137–1150.

Molla, M. A. M., & Ahsan, M. M. (2025). Computers in Human Behavior Reports Artificial intelligence and journalism : A systematic bibliometric and thematic analysis of global research. *Computers in Human Behavior Reports*, *20*(September), 100830. https://doi.org/10.1016/j.chbr.2025.100830

Mu, Y., Jin, M., Song, X., & Aletras, N. (2024). Enhancing Data Quality through Simple De-duplication : Navigating Responsible Computational Social Science Research. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 12477–12492. https://aclanthology.org/2024.emnlp-main.694/

Muchtasor, N. S. (2025). Ethical Pluralism in AI Policy : A Framework for Islamic Integration into Global AI Governance. *Sinergi International Journal of Islamic Studies*, *3*, 191–203. https://journal.sinergi.or.id/index.php/ijis/article/view/897/677

Mutegi, S., Mutegi, D. S., Mamo, Y., Kim, J., Crompton, H., & Mcconnell, M. (2025). Perceptions of STEM education and artificial intelligence : a Twitter (X) sentiment analysis. *International Journal of STEM Education*. https://doi.org/10.1186/s40594-025-00527-5

Nahid, A., 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*, 1–33. https://doi.org/https://doi.org/10.3390/info15110705

Nahm, F. S. (2022). Receiver operating characteristic curve: overview and practical use for clinicians. *Korean Journal of Anesthesiology*, *75*(1), 25–36.

https://doi.org/10.4097/kja.21209

Nariya, M. K., Mills, C. E., Sorger, P. K., & Sokolov, A. (2023). Paired evaluation of machine-learning models characterizes effects of confounders and outliers. *Patterns*, *4*(8), 100791. https://doi.org/10.1016/j.patter.2023.100791

Nayak, S., Savita, & Sharma, Y. K. (2023). A modified Bayesian boosting algorithm with weight-guided optimal feature selection for sentiment analysis. *Decision Analytics Journal*, *8*(June), 100289. https://doi.org/10.1016/j.dajour.2023.100289

Ni, Y., & Ni, W. (2024). A multi-label text sentiment analysis model based on sentiment correlation modeling. *Frontiers in Psychology*, *December*, 1–9. https://doi.org/10.3389/fpsyg.2024.1490796

Nur, M. A., Umar, N., Feng, Z., & Gani, H. (2025). EVALUATION OF INDOBERT AND ROBERTA : PERFORMANCE OF INDONESIAN LANGUAGE TRANSFORMER MODELS IN SENTIMENT. *JIKO (Jurnal Informatika dan Komputer)*, *8*(0173), 121–127. https://doi.org/10.33387/jiko.v8i2.9988

Pfeffer, M. A., Kwok, J., & Wong, W. (2025). Trends and Limitations in Transformer-Based BCI Research. *Applied Sciences*, *Mi*, 1–26. https://doi.org/Pfeffer, M. A., Kwok, J., & Wong, W. (2025). Trends and Limitations in Transformer-Based BCI Research. Applied Sciences, Mi, 1–26.

Poleksić, A., & Martinčić-ipšić, S. (2025). Pretraining and evaluation of BERT models for climate research. *Discover Applied Sciences*, *5*. https://doi.org/https://doi.org/10.1007/s42452-025-07740-5

Pulungan, V., Wicaksono, S. A., & Bachtiar, F. A. (2026). Komparasi Metode Klasterisasi dan Klasifikasi Sentimen untuk Analisis Komentar Sumber Belajar. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *10*(3), 1–10. https://doi.org/https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/16163

Putra, I. A. A., Salmon, & Kusnandar. (2025). Analisis Komentar Youtube

Terhadap Polemik Ijazah Presiden Ke 7 Indonesia Menggunakan Support Vector Machine. *Bulletin of Computer Science Research*, *6*(1), 445–457. https://doi.org/10.47065/bulletincsr.v6i1.883

Putri, N. A. R., & Ardiansyah. (2023). Analisis Sentimen Terhadap Kemajuan Kecerdasan Buatan di Indonesia Menggunakan BERT dan RoBERTa. *Jurnal Sains dan Informatika*, *9*(November), 137–146. https://doi.org/10.34128/jsi.v9i2.649

Qur'an NU Online. (2026a). *Al-Qur'an Surah Al-Baqarah Ayat 30*. https://quran.nu.or.id/al-baqarah/30

Qur'an NU Online. (2026b). *Al-Qur'an Surah Al-Baqarah Ayat 31*. https://quran.nu.or.id/al-baqarah/31

Qur'an NU Online. (2026c). *Al-Qur'an Surah Al-Hujurat Ayat 6*. https://quran.nu.or.id/al-hujurat/6

Qur'an NU Online. (2026d). *Al-Qur'an Surah Al-Isra' Ayat 36*. https://quran.nu.or.id/al-isra/36

Rahman, A. U., Alsenani, Y., Zafar, A., Ullah, K., Rabie, K., & Shongwe, T. (2024). Enhancing heart disease prediction using a self - attention - based transformer model. *Scientific Reports*, 1–13. https://doi.org/10.1038/s41598-024-51184-7

Raza, A., Younas, F., Siddiqui, H. U. R., Rustam, F., Villar, M. G., Alvarado, E. S., & Ashraf, I. (2024). An improved deep convolutional neural network-based YouTube video classification using textual features. *Heliyon*, *10*(16), e35812. https://doi.org/10.1016/j.heliyon.2024.e35812

Rifaldi, D., Famuji, T. S., Okta, B., Miranda, S., & Purma, F. (2025). Evaluasi Sentimen Pengguna ChatGPT Menggunakan Naive Bayes : Tinjauan dari Confusion Matrix dan Classification Report Evaluation ChatGPT User Sentiment using Naive Bayes : A Review of Confusion Matrix and Classification Report. *Jurnal Riset Sistem dan Teknologi Informasi (RESTIA)*, *3*(2), 81–89. https://doi.org/https://doi.org/10.30787/restia.v3i2.1990

Ritonga, I. S., Wanayumini, & Hartama, D. (2025). Sentiment Classification in

Imbalanced Data : Trade-Offs Between. *Jurnal Teknik Informatika*, *18*(2), 303–315. https://doi.org/https://doi.org/10.15408/jti.v18i2.46652

Rønningstad, E., Velldal, E., & Øvrelid, L. (2024). A GPT among Annotators: LLM-based Entity-Level Sentiment Annotation. *LAW 2024 - 18th Linguistic Annotation Workshop, Co-located with EACL 2024 - Proceedings of the Workshop*, 133–139. https://doi.org/https://aclanthology.org/2024.law-1.13

Sahar, R., & Munawaroh, M. (2025). Artificial intelligence in higher education with bibliometric and content analysis for future research agenda. In *Discover Sustainability*. Springer International Publishing. https://doi.org/10.1007/s43621-025-01086-z

Samuel, D., & Øvrelid, L. (2023). Tokenization with Factorized Subword Encoding. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 14143–14161. https://doi.org/10.18653/v1/2023.findings-acl.890

Santin, R. V., & Rezende, S. O. (2025). Systematic review on aspect-based sentiment analysis in cross-domain. *Artificial Intelligence Review*. https://link.springer.com/article/10.1007/s10462-025-11437-x

Sari, A. K., & Abidin, Z. (2026). Comparative Analysis of BERT , RoBERTa and ALBERT Model Performance with Text Data Augmentation in Multilabel Toxic Comment Classification. *Recursive Journal of Informatics*, *4*(1), 40–48. https://doi.org/10.15294/rji.v4i1.25436

Schoene, A. M., Basinas, I., Tongeren, M. Van, & Ananiadou, S. (2022). A Narrative Literature Review of Natural Language Processing Applied to the Occupational Exposome. *International Jornal of Enviromental Reserach and Public Health*. https://doi.org/10.3390/ijerph19148544

Sebők, M., Ring, O., Kis, M. G., Bánóczy, M. B., & Dinnyés, Á. (2024). The geopolitics of vaccine media representation in Orbán's Hungary—an AI-supported sentiment analysis. *Journal of Computational Social Science*, *7*(3), 2897–2920. https://doi.org/10.1007/s42001-024-00325-z

Sevilla, A., Cuevas-Ruiz, P., Rello, L., & Sanz, I. (2025). Artificial Intelligence in Education : Computer-Assisted Learning and AI-guided Tutors. *Italian*

Economic Journal, 1–38. https://doi.org/https://doi.org/10.1007/s40797-025-00354-1

Siino, M., Tinnirello, I., & Cascia, M. La. (2024). Is text preprocessing still worth the time ? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. *Information Systems*, *121*(December 2023), 102342. https://doi.org/10.1016/j.is.2023.102342

Siwinska, P., Lei, J., Castelló, A., Alonso, P., & Ortí, E. S. Q. (2026). Enhancing transformer performance and portability through auto - tuning frameworks. *The Journal of Supercomputing*, 1–22. https://doi.org/https://doi.org/10.1007/s11227-026-08327-6

Šmíd, J., Přibáň, P., Pražák, O., & Král, P. (2024). Czech Dataset for Complex Aspect-Based Sentiment Analysis Tasks. *2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024 - Main Conference Proceedings*, *2014*, 4299–4310. https://doi.org/https://aclanthology.org/2024.lrec-main.384

Song, X., Salcianu, A., Song, Y., Dopson, D., & Zhou, D. (2021). Fast WordPiece Tokenization. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2089–2103. https://doi.org/10.18653/v1/2021.emnlp-main.160

Stefanovitch, N. (2023). Holistic Inter-Annotator Agreement and Corpus Coherence Estimation in a Large-scale Multilingual Annotation Campaign. *Antologi ACL*, 71–86. https://doi.org/10.18653/v1/2023.emnlp-main.6

Straton, N. (2024). Computational model of engagement with stigmatised sentiment: COVID and general vaccine discourse on social media. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *13*(1), 1–18. https://doi.org/10.1007/s13721-024-00456-3

Subakti, A., Murfi, H., & Hariadi, N. (2022). The performance of BERT as data representation of text clustering. *Journal of Big Data*. https://doi.org/10.1186/s40537-022-00564-9

Sun, T., Liu, X., Qiu, X., & Huang, X. (2022). Paradigm Shift in Natural Language Processing. *Machine Intelligence Research*, *19*(June), 169–183.

https://doi.org/10.1007/s11633-022-1331-6

Suprianto, & Ningsih, D. N. W. (2025). Deteksi Risiko Depresi Pada Pengguna Media Sosial Indonesia Menggunakan Indobertweet. *JCSC Journal of Computer Science Cendikia*, *1*(1), 21–28. https://doi.org/https://journal.risetdigital.org/JCSC/article/view/8

Tarumingkeng, R. C. (2024). Natural Language Processing. *Methods*, *245*(November), 53–54. https://doi.org/10.1016/j.ymeth.2025.11.004

Tay, Y. I., Dehghani, M., Bahri, D., & Metzler, D. (2022). Efficient Transformers : A Survey. *ACM Computing Surveys*, *55*(6). https://doi.org/https://doi.org/10.1145/3530811

Tiwari, D., Nagpal, B., Bhati, B. S., Mishra, A., & Kumar, M. (2023). A systematic review of social network sentiment analysis with comparative study of ensemble-based techniques. In *Artificial Intelligence Review* (Vol. 56, Nomor 11). Springer Netherlands. https://doi.org/10.1007/s10462-023-10472-w

Trisna, K. W., & Jie, H. J. (2022). Deep Learning Approach for Aspect-Based Sentiment Classification : A Comparative Review Deep Learning Approach for Aspect-Based Sentiment Classification : A Comparative Review. *Applied Artificial Intelligence*, *36*(1). https://doi.org/10.1080/08839514.2021.2014186

Tucudean, G., Bucos, M., Dragulescu, B., & Caleanu, D. (2024). Natural language processing with transformers : a review. *PeerJ Computer Science*, 1–22. https://doi.org/10.7717/peerj-cs.2222

Tzimiris, S., Nikiforos, S., Nikiforos, M. N., Kermanidis, K. L., & Mouratidis, D. (2025). A Comparative Evaluation of Transformer-Based Language Models for Topic-Based Sentiment Analysis. *Electronics (Switzerland)*. https://www.mdpi.com/2079-9292/14/15/2957

Venkatesan, N., & Arulanand, N. (2022). Implications of Tokenizers in BERT Model for Low-Resource Indian Language. *Journal of Soft Computing Paradigm*, *4*(4), 264–271. https://doi.org/https://doi.org/10.36548/jscp.2022.4.005

Vincent, J., Vincent, J., Santoso, M. I., Pohan, S., Hasani, M. F., Maulina, A., Vincent, J., Ivan, M., Pohan, S., Fikri, M., & Maulina, A. (2025). Extractive Indonesian News Text Summarization using Extractive Indonesian News MBERT , Text Summarization using and RoBERTa Extractive Indonesian News Text Summarization using and. *Procedia Computer Science*, *269*, 1087–1096. https://doi.org/10.1016/j.procs.2025.09.050

Wang, J., Hao, Y., Bai, H., & Yan, L. (2025). Parallel attention recursive generalization transformer for image super-resolution. *Scientific Re*, 1–23. https://doi.org/https://doi.org/10.1038/s41598-025-92377-y

Wicaksono, A. D. A., & Rojali. (2026). Classifcation of Social Media Posts X For Mental Health Symptoms Identification Using NLP Techniques and Transformers Model. *Journal of Computer Science Research*. https://doi.org/10.3844/jcssp.2026.244.259

Widyasari, A. P., & Muljono. (2026). Sentiment Analysis of YouTube Comments on the 2025 DPR RI Demonstration Using Machine Learning. *Journal of Applied Informatics and Computing (JAIC)*, *10*(2). https://doi.org/https://doi.org/10.30871/jaic.v10i2.12254

Xu, Z., Wen, X., Zhong, G., & Fang, Q. (2025). Public perception towards deepfake through topic modelling and sentiment analysis of social media data. *Social Network Analysis and Mining*, *15*(1). https://doi.org/10.1007/s13278-025-01445-8

Yulianti, E., Bhary, N., Abdurrohman, J., Dwitilas, F. W., Nuranti, E. Q., & Husin, H. S. (2024). Named entity recognition on Indonesian legal documents : a dataset and study using transformer-based models. *International Journal of Electrical and Computer Engineering (IJECE)*, *14*(5), 5489–5501. https://doi.org/10.11591/ijece.v14i5.pp5489-5501

Zhang, H., & Shafiq, M. O. (2024). Survey of transformers and towards ensemble learning using transformers for natural language processing. *Journal of Big Data*. https://doi.org/10.1186/s40537-023-00842-0

Zi, X., Liu, F., & Liu, M. (2025). Transformer with Adaptive Sparse Self-Attention for Short-Term Photovoltaic Power Generation Forecasting.

Electronics, 1–24. https://doi.org/https://doi.org/10.3390/electronics14203981

BIODATA PENULIS

Muhammad Mutawakkil Alallah lahir di Bondowoso pada tanggal 21 November 2002. Penulis merupakan anak pertama dari pasangan Bapak Abdus Somad dan Ibu Ismawati. Pendidikan dasar ditempuh di MI Subulus Salam (2009–2014), pendidikan menengah pertama di MTs Subulus Salam (2014–2017), dan pendidikan menengah atas di MA Nurul Jadid (2017–2020). Pada tahun 2020, penulis melanjutkan pendidikan pada Program Studi Teknologi Informasi, Fakultas Teknik, Universitas Nurul Jadid, Paiton, Probolinggo, dan memperoleh gelar Sarjana Komputer (S.Kom.) pada tahun 2024. Pada tahun yang sama, penulis melanjutkan studi pada Program Magister Informatika, Pascasarjana UIN Maulana Malik Ibrahim Malang, dan berhasil menyelesaikan pendidikan magister dalam waktu tiga semester sehingga memperoleh gelar Magister Komputer (M.Kom.).

Minat akademik penulis meliputi bidang kecerdasan buatan (*Artificial Intelligence*), pengolahan bahasa alami (*Natural Language Processing*), pembelajaran mesin (*Machine Learning*), dan analisis data media sosial. Sebagai salah satu syarat untuk memperoleh gelar Magister Komputer (M.Kom.), penulis menyusun tesis berjudul *"Klasifikasi Sentimen Komentar Publik di YouTube tentang Kecerdasan Buatan (AI) dalam Pendidikan Menggunakan BERT dan IndoRoBERTa."* Penelitian ini berfokus pada penerapan model berbasis *transformer* untuk klasifikasi sentimen komentar publik pada platform YouTube terkait pemanfaatan kecerdasan buatan dalam bidang pendidikan.