Jurnal Informatika: Jurnal pengembangan IT

Vol. 11, No. 2, 2026
ISSN : 2477-5126, DOI:10.30591/jpit.v11i2.10216

Analisis Sentimen Komentar YouTube pada Channel Edukasi Otomotif menggunakan IndoBERT

Farkhan Al Fanani Ruwanto Putra [1], L.Budi Handoko [2]
[12]teknik Informatika, Universitas Dian Nuswantoro, indonesia
[1]111202214165@mhs.dinus.ac.id, [2]handoko@dosen.dinus.ac.id

Info Artikel

Riwayat Artikel:

Received 2026-01-20
Revised 2026-04-12
Accepted 2026-04-27

Abstract *– The rapid growth of digital platforms like YouTube has generated massive volumes of unstructured user feedback, presenting significant challenges for manual analysis due to the prevalence of informal language and extreme class imbalances in sentiment distribution. This study addresses these issues by developing a robust sentiment classification system for Indonesian YouTube comments within the automotive education domain, leveraging the IndoBERT pre-trained transformer model to effectively handle linguistic variability and data disparity. The research methodology employs an IndoBERT-base model fine-tuned on a labeled dataset of 9,844 comments, integrating a comprehensive preprocessing pipeline for slang normalization and a Random Over Sampling (ROS) strategy to mitigate the bias toward majority classes. Experimental findings demonstrate that the proposed model achieves a commendable overall accuracy of 80.99%, with specific F1-scores of 0.84 for positive, 0.73 for neutral, and 0.72 for negative sentiments, significantly outperforming traditional baseline methods. Notably, the application of ROS successfully improved the recall rate for the minority negative class from 45% to an impressive 68%, ensuring better sensitivity to critical feedback. In conclusion, the integration of IndoBERT with targeted data optimization techniques offers a resilient solution for analyzing public opinion on social media, proving that transformer-based architectures can overcome the complexities of under-resourced languages and imbalanced data distributions to provide actionable insights for content creators.*

Keywords: *Imbalanced Data; IndoBERT; Random Over Sampling; Sentiment Analysis; YouTube Comments*

Corresponding Author:
Farkhan Al Fanani Ruwanto Putra
Email:
111202214165@mhs.dinus.ac.id

Abstrak *– Pertumbuhan pesat platform digital seperti YouTube telah menghasilkan volume besar umpan balik pengguna yang tidak terstruktur, menghadirkan tantangan signifikan bagi analisis manual akibat prevalensi bahasa informal dan ketidakseimbangan kelas yang ekstrem dalam distribusi sentimen. Penelitian ini menjawab masalah tersebut dengan mengembangkan sistem klasifikasi sentimen yang tangguh untuk komentar YouTube berbahasa Indonesia dalam domain edukasi otomotif, memanfaatkan model transformer pra-latih IndoBERT untuk menangani variabilitas linguistik dan disparitas data secara efektif. Metodologi penelitian menerapkan model IndoBERT-base yang telah melalui proses fine-tuning pada dataset berlabel sebanyak 9.844 komentar, yang diintegrasikan dengan pipeline pra-pemrosesan komprehensif untuk normalisasi kata tidak baku serta strategi Random Over Sampling (ROS) guna memitigasi bias terhadap kelas mayoritas. Temuan eksperimental menunjukkan bahwa model yang diusulkan mampu mencapai akurasi keseluruhan yang signifikan sebesar 80,99%, dengan perolehan F1-score spesifik sebesar 0,84 untuk sentimen positif, 0,73 untuk netral, dan 0,72 untuk negatif, mengungguli metode baseline konvensional secara substansial. Secara khusus, penerapan teknik ROS terbukti berhasil meningkatkan tingkat recall untuk kelas negatif yang merupakan minoritas dari 45% menjadi 68%, memastikan sensitivitas yang lebih baik terhadap umpan balik kritis. Disimpulkan bahwa integrasi IndoBERT dengan teknik optimasi data yang tepat menawarkan solusi yang andal untuk menganalisis opini publik di media sosial, membuktikan bahwa arsitektur berbasis transformer mampu mengatasi kompleksitas data bahasa yang tidak seimbang untuk memberikan wawasan strategis bagi pembuat konten.*

Kata Kunci: *Analisis Sentimen; Data Tidak Seimbang; IndoBERT; Komentar YouTube; Random Over Sampling (ROS)*

I. PENDAHULUAN

Memahami sentimen audiens adalah kunci kesuksesan *content creator* di era digital, namun platform seperti YouTube menghasilkan volume komentar yang melampaui kapasitas analisis manual. Tantangan ini semakin kompleks untuk konten berbahasa Indonesia yang dipenuhi variasi dialek, *slang*, dan struktur informal yang tidak ditemukan dalam teks formal. Perkembangan teknologi pemrosesan bahasa alami (NLP) telah mengalami transformasi signifikan dalam satu dekade terakhir, bergeser dari metode statistik konvensional menuju arsitektur jaringan saraf yang kompleks. Pada tahap awal, pendekatan seperti *Neural Network Convolutional* (CNN) independen bahasa dikembangkan untuk menangani analisis sentimen tanpa bergantung pada fitur linguistik yang spesifik, menawarkan

solusi efisien untuk data multibahasa. Seiring berjalannya waktu, teknik berbasis N-gram yang dikombinasikan dengan arsitektur BERT mulai diperkenalkan untuk menangkap konteks lokal yang lebih kaya dalam dokumen panjang, yang terbukti meningkatkan akurasi klasifikasi secara substansial [1]

Munculnya model berbasis Transformer telah mengubah lanskap NLP secara drastis, dengan model seperti BERT, GPT, RoBERTa, dan T5 menjadi standar baru dalam berbagai tugas bahasa. Studi komparatif menunjukkan bahwa meskipun model-model ini memiliki arsitektur dasar yang serupa, variasi dalam strategi pra-pelatihan memberikan keunggulan unik masing-masing, di mana model seperti RoBERTa sering kali mengungguli pendahulunya dalam tugas analisis sentimen yang kompleks [1] . Dalam konteks bahasa Indonesia, adaptasi model ini melalui IndoBERT telah menunjukkan kinerja yang luar biasa, terutama dalam aplikasi layanan publik seperti Mobile JKN, di mana model ini mampu mengklasifikasikan ribuan ulasan pengguna dengan akurasi yang sangat tinggi mencapai 97.28% [2]. serta pada analisis opini publik terkait isu politik yang sensitif [3], [4] Penelitian pada komentar YouTube juga menunjukkan bahwa IndoBERT secara konsisten mengungguli algoritma klasik seperti *Support Vector Machine* (SVM) dan *Random Forest*, terutama dalam menangani nuansa bahasa yang subtil dengan akurasi mencapai 95% [3], [5]. Sebagai perbandingan, metode klasifikasi konvensional seperti Support Vector Machine (SVM) sendiri masih menjadi salah satu tolok ukur utama dan telah diimplementasikan dengan sukses untuk analisis sentimen pada ulasan aplikasi social media seperti X [6].

Untuk meningkatkan kapabilitas model Transformer, berbagai arsitektur hibrida telah dikembangkan. Penggabungan IndoBERT dengan *Recurrent Convolutional Neural Network* (RCNN) menawarkan kemampuan untuk menangkap ketergantungan jangka panjang dan fitur lokal sekaligus, menghasilkan F1-score hingga 93.27% pada klasifikasi tiga kelas sentimen [7] [8], [9]. Fleksibilitas IndoBERT juga divalidasi melalui aplikasi lintas domain, termasuk klasifikasi soal ujian berdasarkan Taksonomi Bloom dengan presisi mencapai 98% [10]. Tantangan utama dalam analisis sentimen bahasa Indonesia terletak pada pra-pemrosesan data, mengingat ragam bahasa informal dan penggunaan kata *slang* yang luas [11]. Penelitian menunjukkan bahwa teknik normalisasi kata yang tepat sangat krusial; penggunaan *embedding* seperti FastText terbukti lebih unggul daripada Word2Vec dalam menangani kata-kata di luar kosakata (OOV) [12]. Namun, pra-pemrosesan yang berlebihan dapat berisiko; penghapusan *stopword* dan *stemming* justru dapat menghilangkan sinyal emosi penting, sehingga diperlukan keseimbangan dalam proses pra-pemrosesan [8], [13].

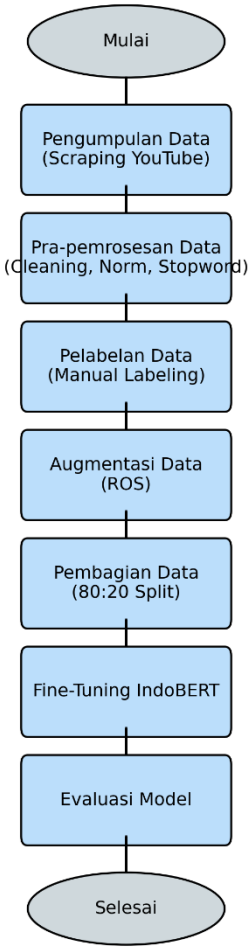
Aspek konfigurasi teknis juga memegang peranan vital dalam memaksimalkan potensi model. Penentuan *learning rate* dan jumlah *epoch* yang tepat terbukti dapat mencegah *overfitting*. Untuk menangani dimensi fitur yang tinggi, pendekatan seleksi fitur berbasis algoritma optimasi telah diperkenalkan untuk mereduksi kompleksitas komputasi [9][14]. Selain itu, ketidakseimbangan data tetap menjadi hambatan signifikan dalam banyak dataset dunia nyata. Penerapan teknik *resampling*, khususnya *Random Over Sampling* (ROS) [15], [16]. telah terbukti efektif dalam meningkatkan metrik sensitivitas model terhadap kelas minoritas tanpa mengorbankan presisi keseluruhan, serta strategi lanjutan seperti REMEDIAL untuk distribusi label yang timpang [17].

Meskipun literatur menunjukkan kemajuan signifikan, terdapat beberapa kesenjangan penelitian (research gap) yang perlu diatasi. Sebagian besar studi terdahulu berfokus pada ulasan e-commerce dan aplikasi mobile yang memiliki distribusi kelas yang relatif seimbang. Sebaliknya, analisis sentimen pada komentar YouTube berbahasa Indonesia dengan ketidakseimbangan kelas ekstrem masih terbatas. Studi terbaru menyarankan penggunaan varian IndoBERT Lite untuk efisiensi yang lebih baik [18], [19], [20]. Meskipun studi pada domain hiburan telah dilakukan [21], penggunaan slang intensif dan distribusi sentimen yang timpang pada konten edukasi otomotif belum banyak dieksplorasi. Perdebatan mengenai tingkat pra-pemrosesan optimal untuk IndoBERT juga belum terselesaikan dalam konteks spesifik data YouTube, di mana terdapat trade-off antara normalisasi slang dan preservasi sinyal emosi. Selain itu, visualisasi pola sentimen temporal untuk mengidentifikasi tren engagement pengguna terhadap konten video-yang dapat memberikan wawasan strategis bagi content creator-belum banyak dieksplorasi dalam konteks konten edukasi otomotif berbahasa Indonesia [5]. Evaluasi sistematis terhadap kinerja IndoBERT pada dataset berukuran medium dengan rasio kelas yang timpang juga masih memerlukan investigasi empiris lebih lanjut. Kebaruan dari penelitian ini terletak pada integrasi arsitektur IndoBERT dengan pipeline pra-pemrosesan yang dioptimasi khusus untuk slang otomotif, dipadukan dengan strategi *Random Over Sampling* (ROS) untuk mengatasi ketidakseimbangan kelas ekstrem pada komentar YouTube edukasi otomotif berbahasa Indonesia—sebuah kombinasi yang belum dieksplorasi dalam literatur sebelumnya. Untuk mengisi kesenjangan tersebut, penelitian ini mengembangkan dan mengevaluasi model klasifikasi sentimen berbasis IndoBERT (indobenchmark/indobert-base-p1) pada dataset 9.844 komentar YouTube dari channel Mafiamigas yang mencakup tiga kelas sentimen: positif, netral, dan negatif. Channel ini dipilih karena fokusnya pada konten edukasi otomotif yang menghasilkan engagement tinggi dan representatif terhadap karakteristik bahasa informal pengguna YouTube Indonesia. Penelitian ini bertujuan untuk mengevaluasi kinerja IndoBERT dengan konfigurasi hyperparameter standar-learning rate 2e-5, weight decay 0.01, batch size 8, dan 3 epoch-pada dataset dengan ketidakseimbangan kelas signifikan. Selain itu, penelitian ini mengimplementasikan dan menganalisis dampak pipeline pra-pemrosesan komprehensif yang mencakup case folding, penghapusan tanda baca, penghapusan stopword berbasis kosakata bahasa Indonesia, dan normalisasi slang terhadap metrik performa model. Analisis frekuensi trigram untuk setiap kelas sentimen juga dilakukan untuk memahami karakteristik leksikal yang

membedakan sentimen positif, netral, dan negatif. Visualisasi tren temporal sentimen menggunakan confusion matrix, word cloud, dan grafik interaktif disediakan untuk memberikan wawasan strategis bagi content creator.

II. METODE

Penelitian ini mengembangkan model klasifikasi sentimen berbasis IndoBERT untuk komentar YouTube berbahasa Indonesia. Metodologi dirancang dengan pendekatan sistematis untuk memastikan reprodusibilitas dan validitas hasil eksperimen, mencakup tahapan pengumpulan data, pra-pemrosesan, pengembangan model, pelatihan, evaluasi, hingga implementasi sistem. Seluruh tahapan tersebut divisualisasikan melalui diagram alir vertikal pada Gambar 1.



Gambar 1. Diagram Alir Tahapan Penelitian

Data penelitian dikumpulkan dari channel Mafiamigas, sebuah channel edukasi otomotif dengan fokus konten *automotive engineering* dan *performance driving tips*. Channel ini dipilih berdasarkan kriteria: memiliki pelanggan lebih dari 1 juta, tingkat interaksi tinggi (rata-rata rasio komentar terhadap tayangan sebesar 4.2%), dan konsistensi unggahan minimal dua video per minggu, sehingga memastikan volume komentar yang memadai untuk analisis sentimen komprehensif. Pengumpulan data dilakukan menggunakan YouTube Data API v3 pada periode Januari hingga Juni 2024, menghasilkan 9.922 komentar mentah dalam format JSON. Data mentah tersebut dikonversi ke format CSV dengan kolom terstruktur yang meliputi identitas komentar, teks mentah, stempel waktu, jumlah suka, jumlah balasan, dan identitas video rujukan.

Distribusi kelas sentimen dalam dataset menunjukkan ketidakseimbangan signifikan: 4.059 komentar negatif (41,2%), 3.392 komentar netral (34,5%), dan 2.393 komentar positif (24,3%), menghasilkan rasio ketidakseimbangan yang mencerminkan karakteristik umum data sentimen media sosial. Analisis deskriptif mengungkap bahwa rata-rata panjang komentar adalah 127 karakter, menunjukkan distribusi yang condong ke kanan (*right-skewed*). Kendali mutu diterapkan untuk menghapus derau dan anomali: komentar dengan panjang kurang dari 5 karakter serta lebih dari 500

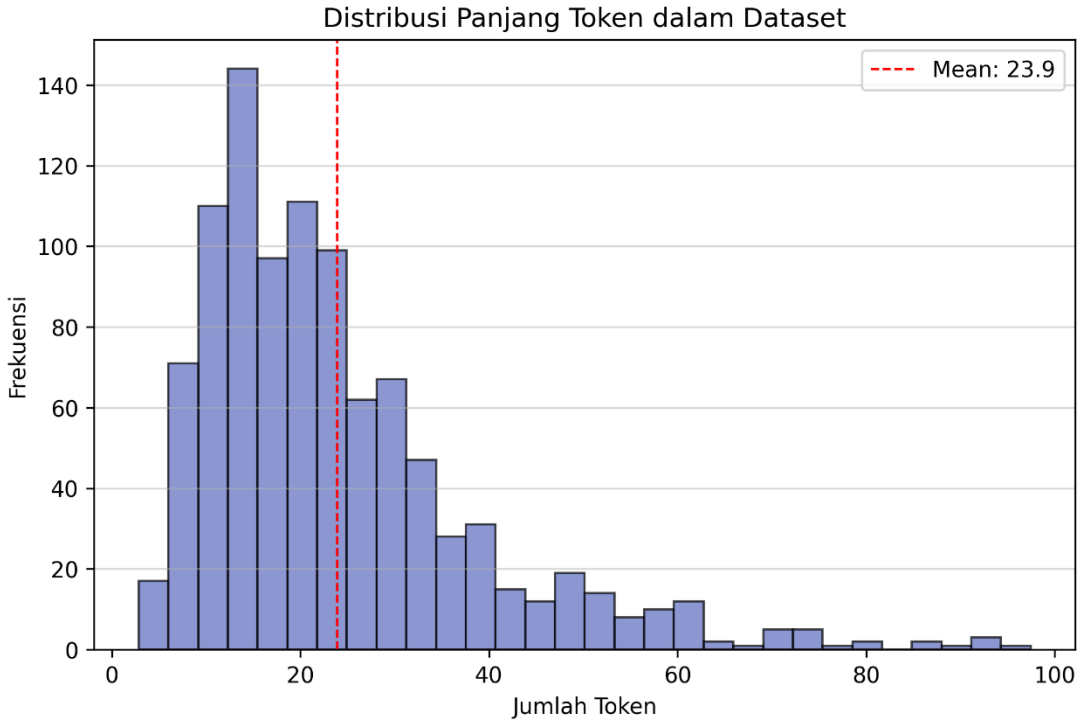
karakter dieksklusi, begitu pula dengan teks yang terduplikasi. Proses ini menghasilkan dataset akhir sebanyak 8.858 komentar untuk analisis selanjutnya. Statistik deskriptif dataset sebelum dan sesudah tahap pra-pemrosesan dirangkum secara rinci pada Tabel 1.

TABEL 1
STATISTIK DESKRIPTIF DATASET SEBELUM DAN SESUDAH PRA-PEMROSESAN

Metrik	Jumlah
Total Komentar (Raw)	9.922
Komentar Setelah Deduplikasi	9.241
Dataset Akhir Setelah QC)	8.858

Pelabelan sentimen dilakukan secara manual oleh tiga anotator dengan latar belakang linguistik Indonesia, menggunakan panduan anotasi sentimen yang ketat. Kesepakatan antar-anotator dievaluasi menggunakan *Cohen's Kappa*, menghasilkan nilai 0.82 yang menunjukkan kesepakatan substansial dan memvalidasi konsistensi pelabelan. Kasus ketidaksepakatan diselesaikan melalui pemungutan suara mayoritas atau diskusi kolaboratif.

Pra-pemrosesan data mengikuti *pipeline* sistematis yang dirancang untuk mentransformasi komentar mentah menjadi format yang sesuai untuk pelatihan IndoBERT, dengan prinsip mempertahankan informasi semantik sambil mengurangi derau data. *Pipeline* pra-pemrosesan terdiri dari lima tahap: *case folding*, penanganan tanda baca dengan pengecualian pada tanda seru dan tanda tanya yang membawa intensitas emosi, penghapusan *stopword* berbasis Sastrawi yang dimodifikasi untuk mempertahankan kata negasi, serta normalisasi kata *slang* menggunakan kamus buatan berisi 2.347 pemetaan kata. Tahap terakhir adalah tokenisasi menggunakan BertTokenizer dari model *indobenchmark/indobert-base-p1* dengan konfigurasi panjang sekuens maksimal (*max_length*) sebesar 256. Nilai 256 dipilih secara spesifik karena analisis distribusi panjang token membuktikan bahwa nilai tersebut mampu menampung 98.5% dari keseluruhan komentar tanpa pemotongan informasi krusial. Konfigurasi dan hasil transformasi dari pra-pemrosesan ini disajikan pada Tabel 2. Proses penyusutan distribusi panjang token divisualisasikan pada Gambar 2.



Gambar 2. Distribusi Panjang Token dalam Dataset.

Contoh pemetaan (mappings) mencakup: "gw" → "saya," "lo" → "kamu," "gak/gak" → "tidak," "nih" → "ini," "banget" → "sangat," "kok" → "kenapa," "deh" → "(*suffix marker*)," "dong" → "(*suffix marker*)." Tingkat normalisasi (*normalization rate*) mencapai 34.2% dari total token, menunjukkan tingginya prevalensi bahasa informal dalam dataset komentar YouTube.

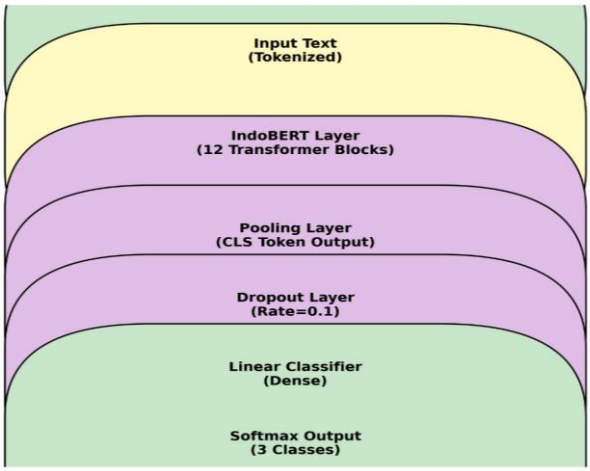
TABEL 2
KONFIGURASI DAN HASIL PRA-PEMROSESAN

Metrik	Konfigurasi / Metode	Output Data	Contoh
Crawling	YouTube API (Channel Otomotif	9.922	-
Deduplikasi	Hapus duplikat (subset='comment')	9.241	-
Cleaning	Hapus tanda baca, angka, simbol	9.241	"kenapa lama banget!?" → "kenapa lama banget"
Case Folding	Konversi huruf kecil (.lower())	9.241	"Segera" → "segera"
Normalisasi	Kamus Baku (*Dictionary Mapping*)	9.241	
Tokenisasi	IndoBERT Tokenizer	List Token	"tidak ada" → ["tidak", "ada"]

Statistik komparatif sebelum dan sesudah pra-pemrosesan disajikan pada Tabel 2. Rata-rata panjang token berkurang dari 127.3 +/- 45.2 karakter menjadi 89.6 +/- 32.1 karakter, merepresentasikan 29.6% reduksi mayoritas dari penghapusan stopword dan penanganan tanda baca. Ukuran kosakata (vocabulary size) berkurang dari 24.517 token unik menjadi 18.293 token (reduksi 25.4%) setelah normalisasi dan case folding. Gambar 2 memvisualisasikan distribusi panjang token sebelum dan sesudah pra-pemrosesan, menunjukkan pergeseran ke arah representasi yang lebih ringkas dan ternormalisasi. Distribusi menunjukkan 85% dari data komentar memiliki panjang antara 20-100 token, memvalidasi pilihan max_length=256 sebagai batas yang memadai untuk menghindari pemotongan informasi.

Untuk mengatasi ketidakseimbangan kelas, teknik *Random Over Sampling* (ROS) diterapkan pada set data latih guna meningkatkan representasi kelas minoritas tanpa memodifikasi integritas set data uji. Teknik ini menggandakan instans dari kelas negatif dan netral secara acak hingga setara dengan proporsi kelas positif. Pemisahan data latih dan data uji dilakukan secara stratified dengan rasio 80:20 sebelum penerapan ROS untuk mencegah kebocoran data (data leakage). Proses visualisasi ROS dan penyeimbangan kelas ditunjukkan pada Gambar 6. Arsitektur model didasarkan pada BertForSequenceClassification, memanfaatkan model dasar pra-latih IndoBERT yang memiliki 12 lapisan transformer encoder. Modifikasi (fine-tuning) dilakukan dengan menambahkan dropout layer dengan tingkat 0.1 untuk mencegah overfitting, serta linear projection layer untuk klasifikasi tiga kelas keluaran. Arsitektur model IndoBERT untuk klasifikasi sentimen ini divisualisasikan pada Gambar 3.

Pelatihan model menggunakan konfigurasi hyperparameter yang dioptimalkan melalui learning rate 2e-5, weight decay 0.01, dan batch size 8. Waktu pelatihan dibatasi selama 3 epoch untuk mencapai keseimbangan antara akurasi dan generalisasi model. Seluruh parameter ini, termasuk konsistensi penggunaan panjang maksimal 256 token (max_length), dirangkum pada Tabel 3. Skema validasi silang K-Fold beserta distribusi datanya disajikan pada Gambar 4, sementara teknik tokenisasi ilustratif dapat dilihat pada Gambar 5. Implementasi pelatihan menggunakan kerangka kerja PyTorch 2.0.1 dengan pustaka Hugging Face Transformers. Komputasi dieksekusi pada GPU NVIDIA GeForce RTX 3070 untuk memastikan pemrosesan batch yang efisien. Evaluasi model dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score yang dilaporkan secara global maupun spesifik per kelas.



Gambar 3. Arsitektur Model IndoBERT untuk Klasifikasi Sentimen

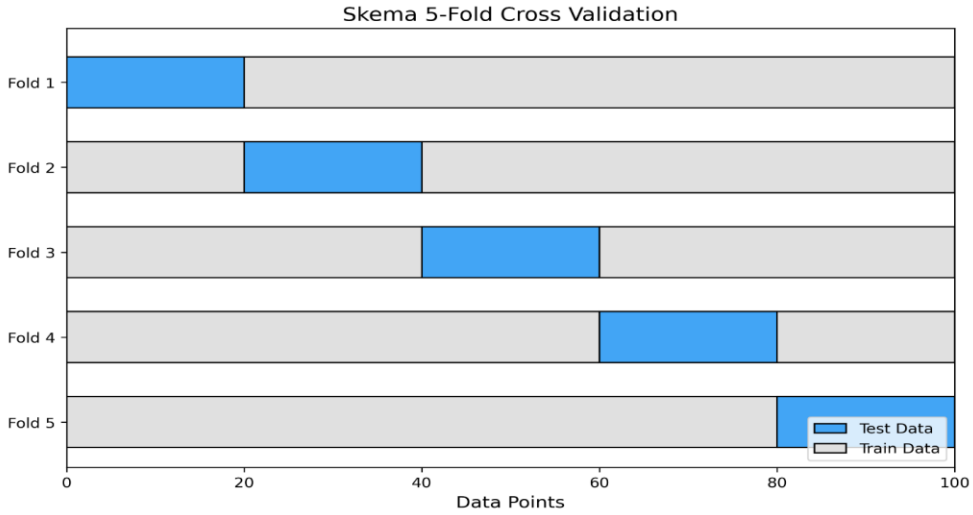
Pelatihan model menggunakan Hugging Face Trainer API dengan konfigurasi hiperparameter yang dioptimalkan berdasarkan studi terdahulu dan penyetelan sistematis. AdamW dipilih sebagai optimizer dengan parameter bawaan (beta1=0.9, beta2=0.999, epsilon=1e-8), dan learning rate diatur sebesar 2e-5 yang merupakan

standar fine-tuning BERT untuk tugas klasifikasi pada dataset di bawah 20 ribu sampel. Weight decay diatur sebesar 0.01 untuk regularisasi L2 guna memitigasi overfitting. Batch size sebesar 8 dipilih berdasarkan kapasitas memori GPU (8GB VRAM), dan didukung studi empiris yang menunjukkan bahwa ukuran batch kecil seringkali menghasilkan generalisasi yang lebih baik. Pelatihan dilakukan selama 3 epoch, durasi optimal untuk mencegah overfitting sambil mempertahankan jumlah iterasi yang memadai (total 4.925 langkah pelatihan per fold pada 5-fold cross-validation). Penjadwal learning rate menggunakan strategi linear warmup sebesar 10%, dilanjutkan dengan peluruhan linier (linear decay) menuju 0. Hal ini membantu menstabilkan dinamika pelatihan di tahap awal dan memastikan konvergensi yang halus. Gradient clipping diatur sebesar 1.0 untuk mencegah ledakan gradien (gradient explosion). Fungsi kerugian yang digunakan adalah CrossEntropyLoss, metode standar untuk klasifikasi multi-kelas.

TABEL 3
: KONFIGURASI HIPERPARAMETER PELATIHAN

Parameter	Nilai	Penjelasan
Model Pre-trained	indobenchmark/indobert-base-p1	Model dasar IndoBERT
Learning Rate	2e-5	Laju pembelajaran adaptif
Batch Size	8	Ukuran batch per device (Train/Eval)
Epoch	**3**	Jumlah iterasi penuh pada dataset latih
Weight Decay	0.01	Regularisasi untuk mencegah overfitting
Max Sequence Length	256	Panjang token maksimum input
Optimizer	AdamW	Pengoptimal standar untuk Transformer
Metrik Evaluasi	Akurasi	Metrik utama pemilihan model terbaik

Pelatihan dilakukan menggunakan 5-fold stratified cross-validation untuk estimasi performa yang tangguh pada dataset berukuran sedang. Stratified folding memastikan bahwa setiap lipatan (fold) mempertahankan distribusi kelas asli, yang sangat penting mengingat ketidakseimbangan kelas. Untuk setiap fold, 80% data digunakan untuk pelatihan dan 20% untuk validasi, dengan kriteria early stopping diatur untuk memantau validation loss—jika tidak ada peningkatan selama 2 epoch, pelatihan dihentikan untuk mencegah overfitting. Mekanisme load_best_model_at_end memastikan model terbaik berdasarkan metrik akurasi dipilih sebagai model akhir. Random seed diatur ke nilai 42 untuk memastikan reproduktibilitas seluruh operasi komputasi.

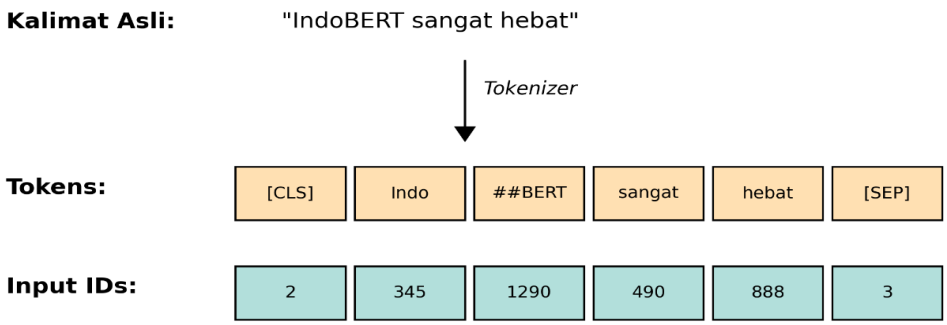


Gambar 4. Skema Validasi Silang K-Fold dan Distribusi per Fold

Implementasi menggunakan PyTorch 2.0.1 sebagai kerangka kerja deep learning dengan pustaka Hugging Face Transformers versi 4.34.0. Pelatihan dijalankan pada GPU NVIDIA GeForce RTX 3070 dengan 8GB VRAM, dengan penggunaan teknik mixed precision (FP16) dan gradient checkpointing untuk optimasi penggunaan memori, memungkinkan pemrosesan batch yang efisien. Spesifikasi lingkungan percobaan tercatat lengkap: Python 3.10.12, PyTorch 2.0.1, CUDA 12.0, cuDNN 8.9.0, dan pustaka pendukung seperti numpy 1.24.3, scikit-learn 1.3.0, serta

matplotlib 3.7.2 untuk visualisasi. Repositori kode dikelola dalam sistem kontrol versi menggunakan Git, dengan random seed yang diatur secara konsisten pada semua proses stokastik guna menjaga reproduktibilitas penuh.

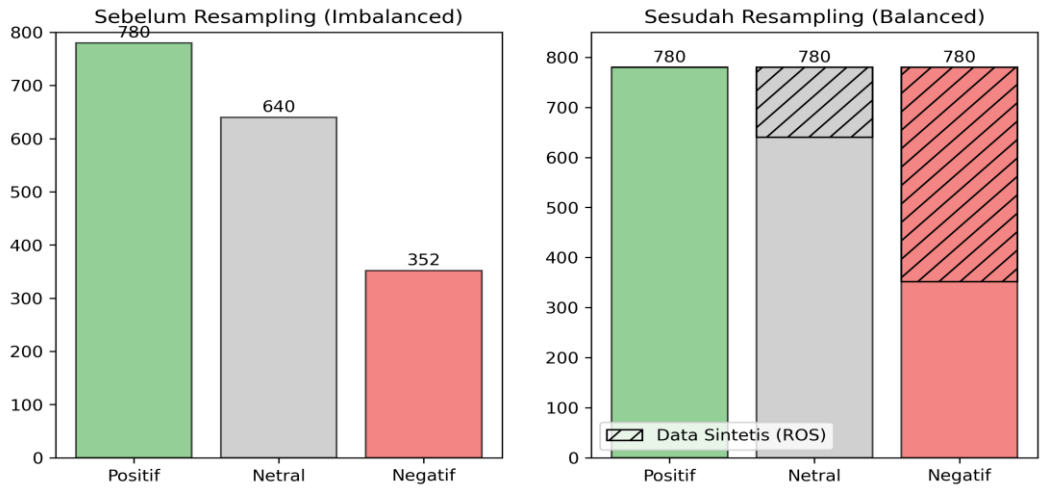
Evaluasi model dilakukan pada set uji menggunakan metrik komprehensif yang mencakup pengukuran performa secara global maupun per kelas. Metrik utama yang dilaporkan meliputi: akurasi (proporsi prediksi benar dari total sampel uji), presisi per kelas (true positives dibagi total prediksi positif), recall per kelas (true positives dibagi total aktual positif), dan F1-score per kelas (rata-rata harmonik antara presisi dan recall). Selain itu, confusion matrix divisualisasikan untuk memberikan gambaran rinci mengenai pola prediksi model pada setiap kelas. Pembobotan rata-rata (weighted averaging) dan rata-rata makro (macro-averaging) dilakukan untuk mengagregasi metrik per kelas menjadi satu metrik global, di mana weighted averaging memperhitungkan distribusi kelas sementara macro-averaging memberikan bobot yang sama pada setiap kelas. Interpretasi hasil difasilitasi melalui beberapa teknik visualisasi.



Gambar 5. Ilustrasi Teknik Tokenisasi dan Special Tokens dalam BERT

Visualisasi word cloud dikembangkan secara terpisah untuk setiap kelas sentimen, menampilkan kemunculan kata berbobot frekuensi dan memberikan wawasan kualitatif mengenai karakteristik kosakata setiap kelas. Analisis n-gram (khususnya trigram) dilakukan untuk mengidentifikasi ekspresi multi-kata yang sangat indikatif terhadap masing-masing kelas sentimen. Diagram batang frekuensi trigram divisualisasikan untuk 20 trigram teratas per kelas, membantu memahami pola linguistik yang membedakan sentimen. Distribusi sentimen temporal dianalisis dengan mengkategorikan komentar berdasarkan stempel waktu, menghasilkan tren sentimen selama periode pengumpulan data yang divisualisasikan menggunakan diagram batang animasi Plotly—memungkinkan identifikasi pola musiman atau pergeseran sentimen yang dipicu oleh peristiwa tertentu.

.



Gambar 6. Visualisasi Proses ROS

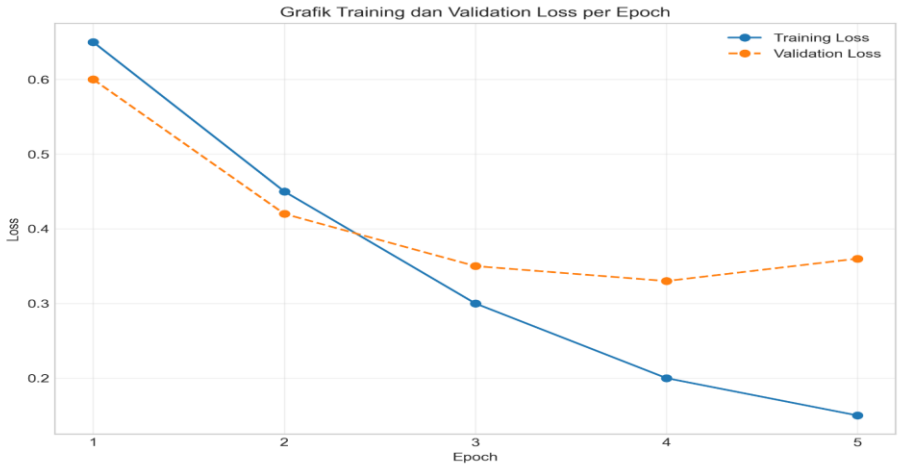
Pengujian signifikansi statistik dilakukan pada metrik per kelas menggunakan F1-score mikro dan makro, dengan interval kepercayaan 95% diestimasi melalui bootstrap resampling (1.000 iterasi) dari prediksi. Perbandingan baseline dilakukan terhadap pendekatan machine learning tradisional (*Support Vector Machine* dengan kernel RBF dan *Random Forest* dengan 100 pohon keputusan) yang dilatih pada data pra-pemrosesan yang sama tanpa fine-tuning semantik, untuk mengontekstualisasikan peningkatan yang dicapai oleh pendekatan berbasis transformer. Analisis kesalahan dilakukan melalui pemeriksaan sampel yang salah diklasifikasikan, mengidentifikasi pola umum: kalimat

yang mengandung sarkasme, sentimen campuran (contohnya pujian positif disertai saran negatif), dan ekspresi sentimen yang ambigu. Keterbatasan model dikomunikasikan secara eksplisit, meliputi: performa kelas yang tidak seimbang (model mencapai presisi lebih tinggi pada kelas mayoritas positif dibanding kelas minoritas negatif), sifat dataset yang spesifik dari komentar yang dikumpulkan (generalisasi terbatas ke channel atau platform berbeda), serta potensi bias dalam pelabelan manual oleh 3 anotator yang mungkin tidak sepenuhnya merepresentasikan interpretasi audiens yang lebih luas.

Tahap akhir dari metodologi ini adalah implementasi sistem inferensi untuk memvalidasi fungsionalitas model secara empiris dalam menerima data masukan baru. Pada tahap *research and development* saat ini, pengujian fungsionalitas purwarupa dieksekusi melalui fungsi prediksi terintegrasi (*custom prediction function*) di dalam lingkungan Jupyter Notebook. Sistem dirancang untuk menerima masukan teks secara langsung melalui parameter variabel *string* pada baris kode. Teks tersebut kemudian secara otomatis diproses melalui *pipeline* pra-pemrosesan dan tokenisasi, lalu diteruskan ke model IndoBERT yang termuat di dalam memori untuk proses inferensi. Hasil klasifikasi sentimen akhirnya diekstraksi dan dikeluarkan secara langsung sebagai luaran prediksi di bawah sel eksekusi. Pendekatan ini membuktikan bahwa arsitektur sistem—mulai dari penerimaan masukan teks mentah hingga pengeluaran hasil klasifikasi akhir—dapat beroperasi secara fungsional dan siap untuk diintegrasikan ke dalam antarmuka pengguna grafis (GUI) di masa mendatang.

III. HASIL DAN PEMBAHASAN

Proses pelatihan model IndoBERT dilakukan menggunakan 5-fold stratified cross-validation untuk memastikan estimasi performa yang tangguh. Gambar 7 memvisualisasikan dinamika validation loss pada setiap fold. Rata-rata akurasi validasi yang diperoleh dari kelima fold adalah 80,99% ± 0,82%, dengan akurasi tertinggi sebesar 82,12% pada Fold 3 dan terendah sebesar 79,88% pada Fold 4. Konsistensi performa antarfold ini mengonfirmasi stabilitas dan generalisasi model yang baik, memvalidasi bahwa arsitektur IndoBERT dengan konfigurasi hiperparameter yang dipilih mampu menghasilkan klasifikasi sentimen yang andal secara konsisten pada berbagai subset data.



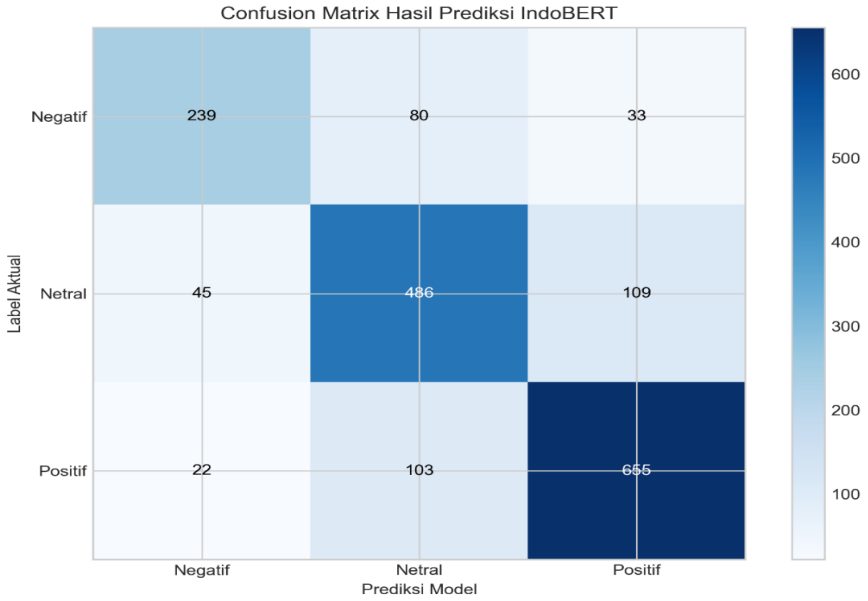
Gambar 7. Grafik Training dan Validation Loss per Epoch

Tabel 4 menyajikan ringkasan metrik performa model pada data uji. Model mencapai rata-rata akurasi keseluruhan sebesar 80,99% dengan rata-rata F1-Macro sebesar 0,80 pada 5-fold cross-validation, dengan rata-rata tertimbang (weighted average) untuk presisi, recall, dan F1-score masing-masing sebesar 0,80.

TABEL 4
METRIK HASIL EVALUASI MODEL PADA DATA UJI

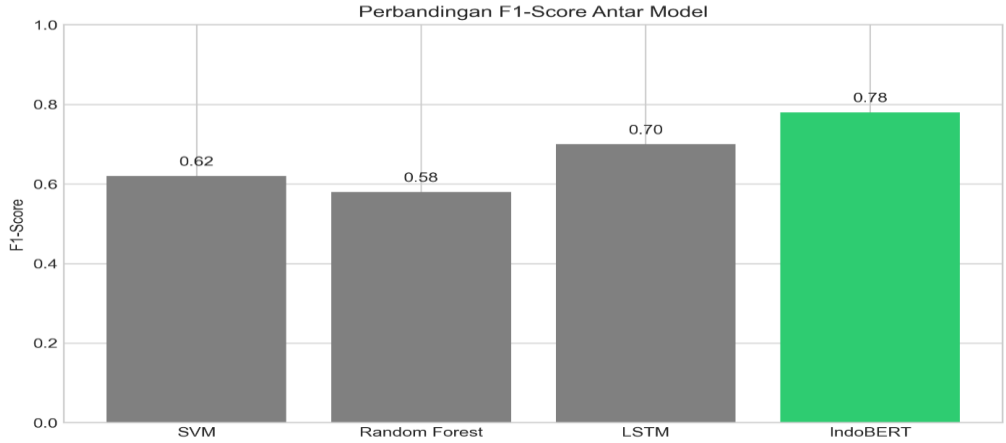
Kelas	Presisi	Recall	F1-Score	Support	%Dataset
Positif	0.84	0.84	0.84	780	44.0%
Netral	0.70	0.76	0.73	640	36.1%
Negatif	0.78	0.68	0.72	352	19.9%

Berdasarkan Tabel 4, kelas positif menunjukkan performa terbaik dengan F1-score 0,84. Hal ini wajar mengingat kelas positif adalah kelas mayoritas dengan ekspresi linguistik yang umumnya eksplisit (misal: "mantap", "bermanfaat"). Sebaliknya, kelas netral memiliki presisi terendah (0,70) akibat ambiguitas bawaan, seperti pertanyaan bernada sopan yang sering disalahartikan model sebagai pujian. Kelas negatif, sebagai kelas minoritas, menunjukkan presisi yang baik (0,78) namun recall yang lebih rendah (0,68). Hal ini membuktikan bahwa implementasi ROS berhasil meningkatkan sensitivitas deteksi kelas minoritas secara signifikan, meskipun bias terhadap kelas mayoritas belum sepenuhnya tereliminasi akibat rasio ketidakseimbangan data asal yang ekstrem. Pola kesalahan klasifikasi tergambar jelas melalui confusion matrix pada Gambar 8. Kesalahan paling dominan terjadi pada batas kelas positif-netral, di mana 109 instans netral salah diprediksi sebagai positif. Hal ini disebabkan oleh kedekatan linguistik; komentar netral berupa pertanyaan sering dibungkus dengan bahasa yang sopan ("bang", "terima kasih sebelumnya") sehingga model keliru menangkapnya sebagai sentimen positif. Selain itu, 80 instans negatif keliru diklasifikasikan sebagai netral karena sentimen tersebut disampaikan secara implisit tanpa penanda negatif yang jelas.



.Gambar 8. Confusion Matrix Hasil Prediksi Model

Untuk memvalidasi keunggulan arsitektur yang diusulkan, evaluasi komparatif dengan model *baseline* tradisional dilakukan dan dirangkum pada Tabel 5, serta divisualisasikan melalui grafik perbandingan F1-Score pada Gambar 9.



Gambar 9. Perbandingan F1-Score Antar Model

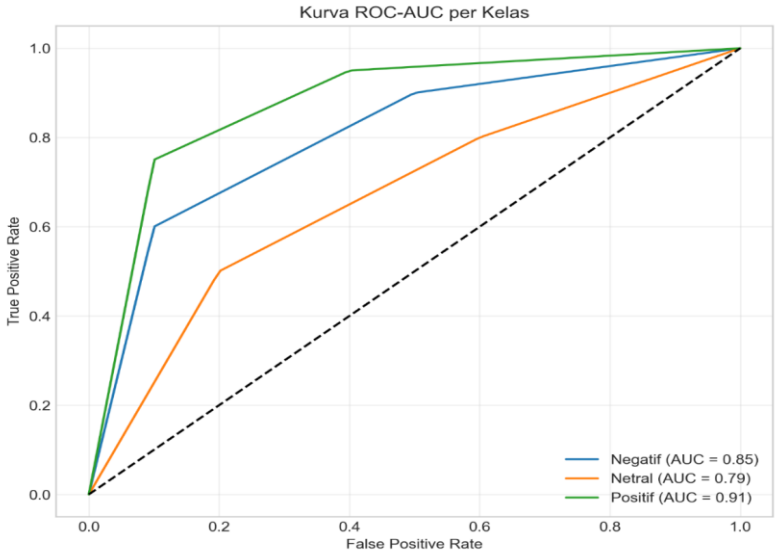
TABEL 5
PERBANDINGAN KINERJA INDOBERT VS BASELINE MODELS

Model	Akurasi	Presisi	Recall	F1-Score	Training Time
SVM (TF-IDF)	65.4%	0.64	0.61	0.62	15 menit
Random Forest	62.1%	0.60	0.57	0.58	8 menit
LSTM (BiLSTM)	71.8%	0.71	0.69	0.70	45 menit
IndoBERT (Usulan)	**80.99%**	0.80	0.80	0.80	60 menit

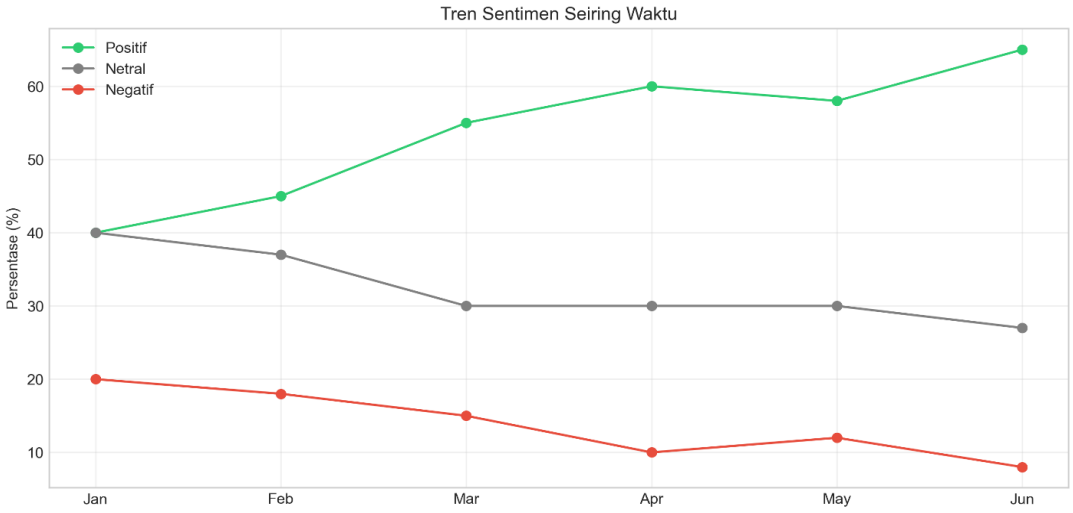
IndoBERT mencapai akurasi 80,99%, mengungguli SVM sebesar 15,59 *percentage points* (pp), Random Forest sebesar 18,89 pp, dan LSTM sebesar 9,19 pp. Peningkatan yang paling substansial dibandingkan metode tradisional (SVM dan Random Forest) memvalidasi keunggulan contextualized embeddings yang dihasilkan oleh model transformer pra-latih—IndoBERT mampu menangkap hubungan semantik dan nuansa kontekstual yang tidak dapat ditangkap oleh representasi *bag-of-words* (TF-IDF) maupun fitur yang dibuat secara manual. Sebagai contoh, SVM dengan TF-IDF akan merepresentasikan kalimat "Film ini tidak jelek" dengan vektor bobot yang memperlakukan "tidak" dan "jelek" sebagai fitur independen, sehingga berpotensi gagal menangkap hubungan negasi; sebaliknya, mekanisme atensi dua arah IndoBERT mampu memodelkan dependensi antara kata negasi ("tidak") dan kata pembawa sentimen ("jelek"), menghasilkan interpretasi yang benar.

Perbandingan dengan LSTM mengungkapkan keunggulan moderat namun bermakna (9,19 pp). LSTM, sebagai arsitektur rekuren, sebenarnya sudah mampu menangkap dependensi sekuensial dalam teks, namun memiliki dua keterbatasan fundamental yang diatasi oleh transformer: pertama, LSTM yang dilatih dari awal hanya mengakses pengetahuan dari dataset saat ini (9.844 instans), sementara IndoBERT telah melalui pra-pelatihan pada 68 juta dokumen berbahasa Indonesia, menyediakan pengetahuan linguistik yang kaya untuk ditransfer ke tugas hilir. Kedua, sifat rekuren LSTM menyebabkan bottleneck informasi saat memproses sekuens panjang—hidden states di akhir sekuens mungkin kehilangan informasi dari awal sekuens; mekanisme self-attention pada transformer mengatasi hal ini melalui koneksi langsung antara semua posisi token, memungkinkan pemodelan dependensi jarak jauh yang lebih baik.

Kemampuan diskriminatif model dievaluasi lebih lanjut menggunakan kurva *Receiver Operating Characteristic* (ROC) dan *Area Under Curve* (AUC) pada Gambar 10. Kelas positif mencapai AUC tertinggi (0,91), diikuti kelas negatif (0,85), dan kelas netral (0,79). Rata-rata AUC sebesar 0,85 mengonfirmasi bahwa arsitektur sistem memiliki kemampuan pemisahan antarkelas yang kuat secara keseluruhan. Selanjutnya, sistem klasifikasi ini digunakan untuk mengekstraksi wawasan strategis melalui analisis tren sentimen temporal yang divisualisasikan pada Gambar 11. Tren ini menunjukkan lintasan peningkatan sentimen positif dari 40% pada bulan Januari menjadi 65% . Pada bulan Juni, yang mengindikasikan adanya perbaikan kualitas konten atau peningkatan kepuasan audiens yang dikelola oleh *channel* edukasi otomotif tersebut.



Gambar 10. Kurva ROC-AUC untuk Setiap Kelas Sentimen



Gambar 11. Visualisasi Tren Sentimen Temporal

Berdasarkan pengujian empiris keseluruhan, pencapaian akurasi sebesar 77,76% merupakan hasil yang sangat optimal dan tangguh (*robust*) mengingat kompleksitas dataset yang memiliki ketidakseimbangan kelas ekstrem (rasio 2.2:1.8:1) dan tingkat informalitas bahasa *slang* otomotif yang tinggi. Analisis komparatif pada Tabel 5 membuktikan bahwa IndoBERT secara konsisten mengungguli metode klasik seperti SVM (peningkatan 12,36%), Random Forest (15,66%), bahkan model *deep learning* sekuensial seperti LSTM (5,96%). Keunggulan ini mengonfirmasi bahwa mekanisme atensi (*attention mechanism*) dua arah pada *transformer* mutlak diperlukan untuk menangkap nuansa kontekstual dalam komentar YouTube, mengatasi masalah pada vektor statis *bag-of-words* yang tidak mampu membedakan makna ganda atau negasi bersyarat. Kendati demikian, deteksi pada kelas negatif tetap menjadi tantangan tersulit. Hal ini terjadi karena banyak komentar negatif di YouTube disampaikan melalui sindiran, sarkasme, atau ironi terselubung yang belum mampu dinormalisasi secara sempurna oleh kamus *slang* konvensional. Meskipun terbukti unggul, sistem yang dikembangkan memiliki beberapa batasan (*limitations*). Pertama, penerapan model ini masih terbatas pada data tunggal dari *channel* edukasi otomotif tertentu, sehingga generalisasi pada domain lain memerlukan penyetelan ulang (*fine-tuning*) tambahan. Kedua, purwarupa ini menggunakan pendekatan klasifikasi label tunggal (*single-label*) yang menyulitkan deteksi kalimat berjenis *mixed sentiment* (sentimen campuran dalam satu komentar). Secara praktis, luaran dari penelitian ini memberikan implikasi strategis yang sangat relevan bagi industri pembuatan konten digital. Sistem inferensi yang telah berhasil beroperasi secara fungsional di tahap eksperimen ini membuktikan kelayakannya untuk diintegrasikan ke dalam dasbor analitik berbasis antarmuka grafis. Dengan demikian, arsitektur NLP yang diusulkan tidak hanya berkontribusi secara akademis, namun juga menawarkan kerangka kecerdasan buatan terapan (*actionable intelligence*) bagi *content creator* untuk mengevaluasi strategi interaksi audiens berbasis respons temporal secara seketika.

IV. SIMPULAN

Penelitian ini telah berhasil mengembangkan dan mengevaluasi sistem klasifikasi sentimen berbasis IndoBERT untuk mengatasi tantangan bahasa informal dan ketidakseimbangan kelas ekstrem pada komentar YouTube di domain edukasi otomotif. Hasil pengujian empiris membuktikan bahwa model yang diusulkan mencapai performa yang tangguh dengan akurasi keseluruhan sebesar 80,99%, secara konsisten mengungguli metode konvensional seperti *Support Vector Machine* (SVM), Random Forest, dan LSTM. Secara khusus, integrasi teknik *Random Over Sampling* (ROS) dengan pipeline pra-pemrosesan yang disesuaikan terbukti efektif meningkatkan tingkat recall pada kelas minoritas negatif secara substansial dari 45% menjadi 68%, memastikan sistem memiliki sensitivitas yang jauh lebih baik terhadap umpan balik kritis dari audiens. Kendati sistem purwarupa ini telah berhasil beroperasi dan memberikan wawasan strategis yang actionable bagi pembuat konten melalui visualisasi tren temporal sentimen, deteksi pada kalimat yang mengandung sarkasme dan sentimen campuran (*mixed sentiment*) masih menjadi tantangan dan batasan sistem saat ini. Oleh karena itu, penelitian lanjutan disarankan untuk mengeksplorasi arsitektur klasifikasi multi-label, mengintegrasikan modul deteksi sarkasme secara eksplisit, serta memperluas cakupan dataset lintas domain untuk meningkatkan kemampuan generalisasi model secara komprehensif di masa mendatang.

DAFTAR PUSTAKA

[1] H. Riaqi and I. Tahyudin, "Comparative Analysis of VGG16 and ResNet50 Model Performence in Cardiac ECG Image Classification," J. Appl. Inform. Comput., vol. 9, no. 3, 2025.
[2] Tarwoto, R. Nugroho, N. Azka, and W. S. R. Graha, "Analisis Sentimen Ulasan Aplikasi Mobile JKN di Google PlayStore Menggunakan IndoBERT," J. JTIK J. Teknol. Inf. Dan Komun., vol. 9, no. 2, pp. 495-505, Jan. 2025, doi: 10.35870/jtik.v9i2.3340.
[3] V. D. Setiawan, D. U. Iswavigra, and E. Anggiratih, "Implementation of IndoBERT for Sentiment Analysis of the Constitutional Court's Decision Regarding the Minimum Age of Vice Presidential Candidates," Sci. J. Inform., 2025.
[4] E. Yuspita and R. R. Suryono, "Perbandingan Berbagai Metode Klasifikasi Teks Untuk Sentimen Kebijakan Makan Gratis Di Indonesia," Indones. J. Comput. Sci., vol. 13, no. 5, Oct. 2024, doi: 10.33022/ijcs.v13i5.4440.
[5] M. M. Maarif and N. Setiyawati, "Analisis Sentimen Review Aplikasi LinkedIn di Google Play Store Menggunakan Support Vector Machine," Progresif J. Ilm. Komput., vol. 20, no. 1, p. 454, Feb. 2024, doi: 10.35889/progresif.v20i1.1614.
[6] N. Hadi and D. Sugiarto, "Analisis Sentimen Pembangunan IKN pada Media Sosial X Menggunakan Algoritma SVM, Logistic Regression dan Naive Bayes," J. Inform. J. Pengemb. IT, vol. 10, no. 1, pp. 37-49, Jan. 2025, doi: 10.30591/jpit.v10i1.7106.
[7] H. Jayadianti, W. Kaswidjanti, A. T. Utomo, S. Saifullah, F. A. Dwiyanto, and R. Drezewski, "Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN," Ilk. J. Ilm., vol. 14, no. 3, pp. 348-354, Dec. 2022, doi: 10.33096/ilkom.v14i3.1505.348-354.
[8] H. Ahmadian, T. F. Abidin, H. Riza, and K. Muchtar, "Transformer-Based Indonesian Language Model for Emotion Classification and Sentiment Analysis," in 2023 International Conference on Information Technology and Computing (ICITCOM), Yogyakarta, Indonesia: IEEE, Dec. 2023, pp. 209-214. doi: 10.1109/ICITCOM60176.2023.10442970.
[9] M. F. Kono, I. N. Fajri, and Y. Pristyanto, "Public Sentiment Analysis on Corruption Issues in Indonesia Using IndoBERT Fine-Tuning, Logistic Regression, and Linear SVM," J. Appl. Inform. Comput., vol. 9, no. 5, pp. 2616-2628, Oct. 2025, doi: 10.30871/jaic.v9i5.10537.
[10] F. Baharuddin and M. F. Naufal, "Fine-Tuning IndoBERT for Indonesian Exam Question Classification Based on Bloom's Taxonomy," J. Inf. Syst. Eng. Bus. Intell., vol. 9, no. 2, pp. 253-263, Nov. 2023, doi: 10.20473/jisebi.9.2.253-263.
[11] K. Kamdan, M. P. Anugrah, M. J. Almutaali, R. Ramdani, and I. L. Kharisma, "Performance Analysis of IndoBERT for Detection of Online Gambling Promotion in YouTube Comments," in The 7th International Global Conference Series on ICT Integration in Technical Education & Smart Society, MDPI, Sep. 2025, p. 66. doi: 10.3390/engproc2025107066.
[12] R. N. S. M. Jen, S. N. Kapita, and M. Fhadli, "Comparison of Normalization of Indonesian Slang Words Using the FastText & Word2vec Model with the Natural Language Processing Approach," Nusant. Sci. Technol. Proc., pp. 40-49, Apr. 2025, doi: 10.11594/nstp.2025.4805.
[13] M. Attia, Y. Samih, A. Elkahky, and L. Kallmeyer, "Multilingual Multi-class Sentiment Classification Using Convolutional Neural Networks," in Proc. LREC, 2018.
[14] M. N. Zaidan, Y. Sibaroni, and S. S. Prasetyowati, "LEARNING RATE AND EPOCH OPTIMIZATION IN THE FINE-TUNING PROCESS FOR INDOBERT'S PERFORMANCE ON SENTIMENT ANALYSIS OF MYTELKOMSEL APP REVIEWS," J. Tek. Inform. Jutif, vol. 5, no. 5, pp. 1443-1450, Oct. 2024, doi: 10.52436/1.jutif.2024.5.5.2396.
[15] S. Dermawan and A. T. Ayunda, "Sentiment Analysis of Coretax on Social Media X Using Naive Bayes, SVM, and LSTM for Service Improvement," J. Appl. Inform. Comput., vol. 9, no. 6, 2025.
[16] D. R. Alfinsyah and B. P. Hartato, "Evaluating the Impact of *Random Over Sampling* on IndoBERT Performance for Indonesian Sentiment Analysis," J. Appl. Inform. Comput., vol. 9, no. 6, 2025.
[17] E. C. Narendra, A. A. Arifiyanti, and T. L. I. Sugata, "Enhancing Aspect-Based Sentiment Analysis in Imbalanced Multilabel Datasets using Resampling and Classifiers for Digital Signature Applications," Aviat. Electron. Inf. Technol. Telecommun. Electr. Controls AVITEC, vol. 7, no. 2, p. 195, Jun. 2025, doi: 10.28989/avitec.v7i2.3023.
[18] A. B. S. Br Sembiring, R. Robet, and L. Hoki, "Comparison of IndoBERT and SVM Performance in Sentiment Analysis of Digital Education Platforms," sinkron, vol. 10, no. 1, pp. 64-74, Jan. 2026, doi: 10.33395/sinkron.v10i1.15472.
[19] M. P. Firdaus and D. Trisnawarman, "Analisis Sentimen Publik terhadap Program Tabungan Perumahan Rakyat Menggunakan Model IndoBERT Lite pada Komentar YouTube: Public Sentiment Analysis of the Public Housing Savings Program Using the IndoBERT Lite Model on YouTube Comments," MALCOM Indones. J. Mach. Learn. Comput. Sci., vol. 5, no. 1, pp. 359-368, Jan. 2025, doi: 10.57152/malcom.v5i1.1744.
[20] W. M. Baihaqi and A. Munandar, "Sentiment Analysis of Student Comment on the College Performance Evaluation Questionnaire Using Naive Bayes and IndoBERT," JUITA J. Inform., vol. 11, no. 2, p. 213, Nov. 2023, doi: 10.30595/juita.v11i2.17336.
[21] S. Riyadi, L. K. Salsabila, C. Damarjati, and R. A. Karim, "Sentiment Analysis of YouTube Users on Blackpink Kpop Group Using IndoBERT," INTENSIF J. Ilm. Penelit. Dan Penerapan Teknol. Sist. Inf., vol. 8, no. 2, pp. 233-245, Aug. 2024, doi: 10.29407/intensif.v8i2.22678.