Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 1, February 2026, Page. 286-306
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5402

Aspect-Based Sentiment Analysis of Access by KAI Application Reviews Using IndoBERT for Multi-Label Classification Tasks

Hilda Nur Alfiana[1], Afrizal Doewes*[2], Bambang Widoyono[3]

[1,2,3]Informatics, Universitas Sebelas Maret, Indonesia

Email: [2]afrizal.doewes@staff.uns.ac.id

Received : Oct 30, 2025; Revised : Nov 21, 2025; Accepted : Nov 21, 2025; Published : Feb 15, 2026

Abstract

Ratings and reviews on mobile applications provide valuable insights into user experience and satisfaction with app features and services. However, ratings are subjective and often inconsistent with the content of the reviews. Therefore, a more in-depth analysis of the review content is necessary to identify evaluation points accurately. This study aims to evaluate the performance of IndoBERT in Aspect-Based Sentiment Analysis (ABSA) on Access by KAI application reviews. Data were collected by scraping user reviews from the Google Play Store, then annotated using a hybrid labeling approach. The resulting dataset was used to fine-tune the IndoBERT model across three ABSA tasks: aspect classification, sentiment classification for each aspect, and joint aspect-sentiment classification. We also benchmarked the model against baseline models to demonstrate its effectiveness. The results show that IndoBERT achieved the best performance across all tasks, specifically aspect classification (accuracy 0.928, F1-score 0.785), sentiment classification (accuracy 0.928, F1-score 0.752), and joint aspect-sentiment classification (accuracy 0.962, F1-score 0.549). Overall, IndoBERT successfully outperformed SVM and XGBoost with TF-IDF, BiLSTM with pre-trained IndoBERT embeddings, mBERT, and XLM-R. This study contributes a new dataset that provides resources for further research and development in Indonesian Natural Language Processing (NLP). These findings also highlight the advantages of a monolingual model trained specifically on Indonesian-language data.

Keywords : *Access by KAI, Aspect-based sentiment analysis, IndoBERT, Multi-label classification, Sentiment analysis*

1. PENDAHULUAN

Jumlah penumpang kereta api di Indonesia pada tahun 2024 mencapai lebih dari 500 ribu orang dan menunjukkan peningkatan signifikan sebesar 35% dibandingkan tahun sebelumnya [1]. Hal ini menunjukkan adanya tren peningkatan minat dan kepercayaan masyarakat terhadap layanan transportasi kereta api. Aplikasi Access by KAI merupakan platform yang dikembangkan untuk memudahkan pengguna dalam bertransaksi tiket kereta api, serta mengakses informasi dan layanan terkait secara daring. Seorang pengguna dapat memberikan *rating* dan ulasan yang mencerminkan pengalaman mereka dalam menggunakan aplikasi pada platform Google Play Store. Penilaian ini memiliki peran penting sebagai sumber umpan balik yang memberikan wawasan berharga mengenai kualitas aplikasi. Namun, skor *rating* yang diberikan oleh pengguna seringkali bersifat subjektif dan tidak mencerminkan isi ulasan [2]. Dengan demikian, perlu dilakukan analisis lebih mendalam terhadap isi ulasan untuk memahami penilaian pengguna serta mengidentifikasi permasalahan mereka. Namun, banyaknya teks ulasan yang ada dengan berbagai konteks dapat menyebabkan pengembang aplikasi kesulitan untuk membaca data dan mengidentifikasi aspek evaluasi. Hal ini juga dapat menyulitkan calon pengguna dalam memahami ulasan pengguna lain ketika mempertimbangkan untuk menggunakan layanan.

Analisis sentimen merupakan sebuah teknik dalam *Natural Language Processing* (NLP) yang umumnya digunakan untuk mengidentifikasi sentimen pada ulasan pengguna. Teknik ini umumnya

Vol. 7, No. 1, February 2026, Page. 286-306
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5402

digunakan pada bisnis untuk mengukur reputasi, memahami kebutuhan pelanggan, sehingga memfasilitasi pengambilan keputusan dalam pengembangan produk [3]. Pada kenyataannya, satu ulasan pengguna seringkali berisi beberapa polaritas opini dengan sentimen yang berbeda-beda terhadap topik atau aspek yang berbeda pula. Namun, model klasifikasi sentimen biasa tidak dapat menangkap sentimen majemuk dan hanya akan mengambil salah satu jenis sentimen saja untuk mewakili keseluruhan isi dalam sebuah ulasan [4]. Oleh karena itu, *Aspect-Based Sentiment Analysis* (ABSA) dapat digunakan untuk mengekstraksi informasi yang lebih terperinci dengan mengidentifikasi aspek yang dibahas dan sentimen dari setiap aspek tersebut.

Penelitian ABSA dalam bahasa Indonesia pada ulasan aplikasi telah dilakukan dengan berbagai pendekatan. Beberapa penelitian sebelumnya menerapkan model berbasis *machine learning*, yaitu *Support Vector Machine* (SVM) dengan ekstraksi fitur TF-IDF untuk melakukan klasifikasi aspek dan sentimen [5], [6]. Pendekatan serupa dilakukan oleh [7], [8] dengan membandingkan performa model SVM dan *Naive Bayes Classifier* (NBC), dan menunjukkan bahwa SVM memperoleh performa yang lebih baik. Selain itu, penelitian lain menerapkan algoritma Random Forest, Decision Tree, dan *Extreme Gradient Boosting* (XGBoost) dengan fitur TF-IDF [9], serta pendekatan *ensemble machine learning* untuk meningkatkan hasil klasifikasi [10]. Beberapa studi juga mengadopsi model berbasis *deep learning* seperti *Bidirectional Long Short-Term Memory* (BiLSTM) [11], *Convolutional Neural Network* (CNN) [12], [13], dan *Multilayer Perceptron* (MLP) dengan *word embeddings* [14].

Tabel 1. Penelitian Terkait

Referensi	Dataset	Model	Hasil
Mustakim dan Priyanta [15]	Ulasan aplikasi KAI Access	SVM, NBC	SVM memperoleh performa terbaik (*accuracy* 91,63%, *F1-score* 75,55%)
Maghfiroh et al. [16]	Ulasan aplikasi Access by KAI	SVM	SVM memperoleh hasil *accuracy* 0.47 dan *F1-score* 0.32
Faradita dan Sibaroni [17]	Ulasan aplikasi KAI Access	CNN-LSTM	CNN-LSTM memperoleh performa terbaik (*accuracy* 87.87%)
Pranata et al. [18]	Ulasan pelanggan PLN	IndoBERT, SVM, LSTM	IndoBERT memperoleh performa terbaik (*F1-score* 0.756)
Yulianti dan Nissa [19]	Ulasan Hotel	IndoBERT	IndoBERT memperoleh performa terbaik (*accuracy* 0.958, *F1-score* 0.884)

Beberapa penelitian telah menerapkan ABSA pada ulasan aplikasi Access by KAI. Ringkasan penelitian yang paling terkait dengan penelitian ini disajikan pada Tabel 1. Penelitian yang dilakukan oleh [15] menerapkan model SVM dan NBC dengan ekstraksi fitur TF-IDF. Namun, data yang digunakan pada penelitian tersebut berada pada rentang waktu 2018 hingga 2020 yang merepresentasikan versi aplikasi terdahulu yaitu KAI Access, sehingga belum mencerminkan kondisi dan fitur aplikasi terbaru Access by KAI. Penelitian serupa oleh [16] juga menggunakan model SVM dengan TF-IDF pada klasifikasi sentimen. Penelitian tersebut masih menghadapi masalah pada keterbatasan jumlah data serta perhitungan evaluasi model yang masih menggunakan *exact-match accuracy*, di mana prediksi dianggap benar hanya jika semua label pada satu data multilabel berhasil diprediksi secara tepat. Dengan demikian, kesalahan prediksi pada salah satu aspek saja akan menyebabkan seluruh prediksi pada data tersebut dianggap salah. Selain itu, kedua penelitian tersebut masih menggunakan pendekatan *machine learning* konvensional dan fitur statis TF-IDF, yang memiliki keterbatasan dalam menangkap konteks semantik dan relasi antar kata, terutama dalam kalimat bahasa

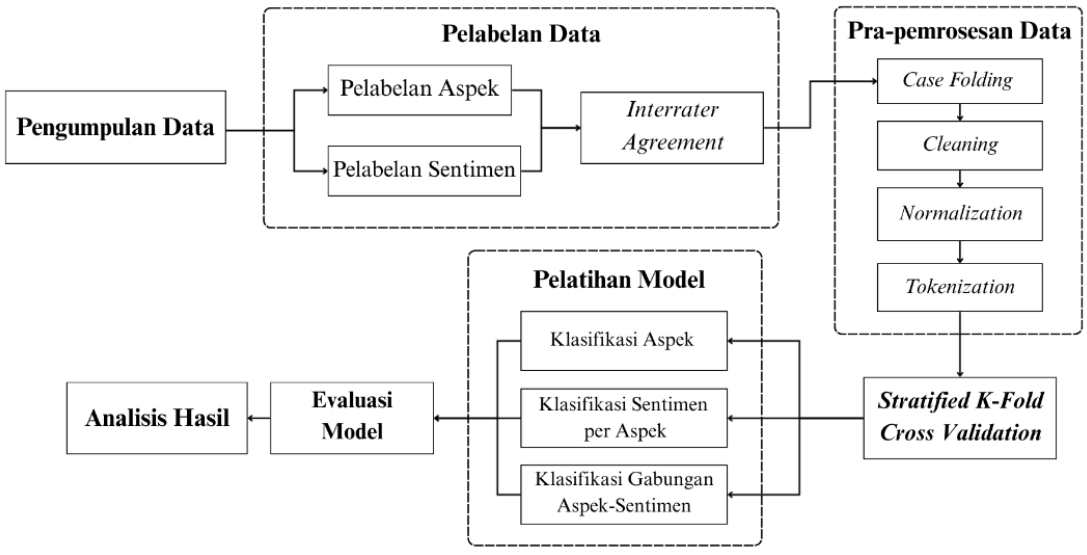
Vol. 7, No. 1, February 2026, Page. 286-306
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5402

Indonesia yang kompleks. Penelitian ABSA pada aplikasi KAI Access juga telah dilakukan menggunakan model CNN–LSTM [17]. Namun, ekstraksi fitur berbasis CNN pada penelitian tersebut masih kesulitan dalam membedakan elemen teks yang memiliki kemiripan pola bahasa, karena representasi yang dihasilkan bersifat lokal dan tidak kontekstual. Penelitian tersebut juga belum menerapkan klasifikasi multi-label, sehingga setiap ulasan hanya dikaitkan dengan satu kategori aspek.

IndoBERT merupakan model berbasis *Bidirectional Encoder Representations from Transformers* (BERT) yang telah dilatih secara khusus menggunakan kumpulan data berbahasa Indonesia yang komprehensif, sehingga efektif dalam menangkap semantik tingkat kalimat [20]. IndoBERT mampu menghasilkan *contextual embedding* yang merepresentasikan makna kata secara dinamis sesuai dengan konteks kalimatnya. Pemilihan model berbasis BERT didasarkan pada kemampuannya dalam menangani struktur teks bahasa Indonesia yang kompleks [21]. Hal ini dibuktikan pada penelitian [18] yang menerapkan ABSA pada data ulasan pelanggan PLN, dan menunjukkan bahwa performa model IndoBERT memperoleh hasil terbaik dan mengungguli model LSTM dan SVM dengan TF-IDF. Penelitian serupa oleh [19] juga membuktikan bahwa model IndoBERT berhasil mengungguli model berbasis BERT lainnya seperti mBERT, distilBERT, serta pendekatan *feature-based* IndoBERT.

Penelitian ini bertujuan untuk menerapkan analisis sentimen berbasis aspek pada ulasan aplikasi Access by KAI menggunakan model IndoBERT. Kinerja model IndoBERT dievaluasi pada tiga tugas klasifikasi, yaitu klasifikasi aspek, klasifikasi sentimen per aspek, dan klasifikasi gabungan aspek-sentimen. Penelitian ini juga melakukan perbandingan model dengan menguji sejumlah model pembanding untuk membuktikan efektivitas model yang diusulkan. Berbeda dari penelitian [16] yang menggunakan *exact-match accuracy*, penelitian ini menerapkan *label-based accuracy* untuk mengevaluasi kinerja model pada setiap label secara individual sehingga menghasilkan penilaian yang lebih komprehensif. Selain itu, penelitian ini juga memberikan kontribusi dengan mengembangkan dataset ABSA terbaru dari ulasan aplikasi Access by KAI yang dikumpulkan dari Google Play Store.

2. METODE PENELITIAN



Gambar 1. Diagram Metode Penelitian

Penelitian ini dilakukan dengan sejumlah tahapan sebagaimana ditunjukkan pada Gambar 1. Tahapan penelitian dibagi menjadi 7 tahapan utama, yaitu (1) Pengumpulan Data, (2) Pelabelan Data,

(3) Pra-pemrosesan Data, (4) Stratified K-Fold Cross Validation, (5) Pelatihan Model, (6) Evaluasi Model, dan (7) Analisis Hasil.

2.1. Pengumpulan Data

Pada penelitian ini, data diperoleh dari ulasan aplikasi Access by KAI yang tersedia di platform Google Play Store yang dikumpulkan dengan metode *scraping* melalui *library google-play-scraper* dengan bahasa pemrograman Python. Data tersebut diambil pada rentang waktu satu tahun, yaitu dari bulan Februari 2024 hingga Februari 2025 dengan jumlah 4231 data.

2.2. Pelabelan Data

Pada tahap ini, data yang telah dikumpulkan selanjutnya diberikan label dengan dua kategori utama, yaitu kategori aspek dan kategori sentimen. Kategori aspek terdiri atas "*Performa*", "*Tampilan*", "*Tiket*", "*Pembayaran*", dan "*Akun*", yang ditentukan berdasarkan hasil observasi pada dataset serta mengacu pada kombinasi aspek-aspek yang digunakan pada penelitian terdahulu [10], [17], [22], [23]. Setiap ulasan dapat memiliki lebih dari satu label aspek, dan setiap aspek yang teridentifikasi diberikan salah satu label sentimen. Adapun kategori sentimen meliputi "*Positif*", "*Negatif*", "*Netral*", dan "*Konflik*". Label sentimen "*Netral*" diberikan jika tidak terdapat indikasi emosi yang kuat baik positif maupun negatif terhadap aspek terkait. Sementara itu, label sentimen "*Konflik*" diberikan ketika terdapat sentimen yang saling bertentangan baik positif maupun negatif terhadap aspek yang sama [24]. Hanya data berlabel sentimen "*Positif*" dan "*Negatif*" yang digunakan pada penelitian ini. Hal ini disebabkan oleh jumlah data berlabel "*Netral*" dan "*Konflik*" yang sedikit, dengan persentase masing-masing label hanya sebesar 0.88% dan 1.25% dari keseluruhan label, sehingga dinilai kurang representatif untuk proses pelatihan model.

Proses pelabelan data melibatkan kombinasi pelabelan otomatis dan manual karena keterbatasan sumber daya. Pelabelan otomatis dilakukan menggunakan API OpenAI dengan model *GPT-4o-mini* pada keseluruhan jumlah data. Sementara itu, pelabelan manual dilakukan oleh seorang pakar yang berpengalaman pada bidang bahasa Indonesia sebanyak 1000 data sebagai acuan utama, dan oleh penulis sebanyak 600 data. Selanjutnya, untuk memastikan validitas dan konsistensi dari pelabelan otomatis dan pelabelan manual penulis terhadap pakar, dilakukan perhitungan *interrater agreement* menggunakan *Cohen's Kappa* pada data yang tumpang tindih antar pelabel. *Interrater agreement* digunakan untuk mengukur tingkat kesepakatan antar pelabel untuk memastikan anotasi data tidak bersifat subjektif. Sebagai pelengkap, *percent agreement* juga dihitung untuk memberikan gambaran proporsional mengenai kesesuaian anotasi antar pelabel. Perhitungan tingkat kesepakatan ini dilakukan pada label aspek dan label sentimen secara terpisah antar pasangan pelabel.

Cohen's Kappa merupakan sebuah indikator yang digunakan untuk mengukur tingkat kesepakatan atau *agreement* antara dua pelabel pada data klasifikasi. Perhitungan kesepakatan dilakukan dengan mempertimbangkan kemungkinan bahwa kesepakatan tersebut terjadi secara kebetulan [25]. Formula Cohen's Kappa ditampilkan pada Persamaan (1).

$$\kappa = \frac{P_o - P_e}{1 - P_e} \qquad (1)$$

P_o merupakan *observed agreement* atau proporsi data sepakat antar pelabel, dan P_e merupakan *expected agreement* atau proporsi kesepakatan yang diharapkan terjadi secara kebetulan. Penafsiran tingkat kesepakatan nilai kappa ditunjukkan pada Tabel 2. Nilai κ berkisar antara -1 hingga 1, yang merepresentasikan keadaan dari *poor* hingga *almost perfect agreement*.

Tabel 2. Interpretasi Nilai Kappa [26]

Nilai Kappa	Tingkat Kesepakatan
< 0.00	*Poor*
0.00 – 0.20	*Slight*
0.21 – 0.40	*Fair*
0.41 – 0.60	*Moderate*
0.61 – 0.80	*Substantial*
0.81 – 1.00	*Almost Perfect*

2.3. Pra-pemrosesan Data

Tahap pra-pemrosesan data mencakup proses mengubah bentuk data tidak terstruktur menjadi lebih terstruktur. Penelitian ini melakukan beberapa tahapan pra-pemrosesan teks, yaitu sebagai berikut.

a. *Case folding,* yaitu proses mengubah seluruh karakter huruf pada dataset menjadi bentuk *lowercase.* Hal ini dilakukan untuk menyetarakan semua kata menjadi bentuk huruf yang sama [27].
b. *Cleaning,* yaitu proses menghilangkan karakter atau elemen tidak diperlukan yang tidak memiliki arti dan dapat menyebabkan noise pada data teks [27]. Pada penelitian ini, dilakukan penghapusan karakter nonalfanumerik selain tanda tanya (?), tanda seru (!), tanda titik (.), dan tanda koma (,).
c. *Normalization,* yaitu proses mengubah variasi kata atau kalimat menjadi bentuk standar, misalnya mengganti bentuk kata singkatan atau kata tidak baku dengan bentuk yang umum atau baku [27]. Penelitian ini menggunakan daftar kata slang bahasa Indonesia yang diperoleh dari repositori Github *open-source* [28].
d. *Tokenization,* yaitu proses pemotongan string input menjadi satuan kata. Penelitian ini menggunakan IndoBERT *Tokenizer* dari *Hugging Face Transformers* versi *indobert-base-p2*. *Hugging Face* merupakan platform yang menyediakan berbagai model pra-latih yang dapat dimanfaatkan untuk penerapan dan pengembangan teknik NLP. *Tokenizer* ini bekerja dengan melakukan *subword tokenization*, yaitu memecah kata menjadi sub-unit kata yang lebih kecil [29].

2.4. *Stratified K-Fold Cross Validation*

Sebelum melakukan pelatihan model, dilakukan pemisahan data menggunakan metode *stratified k-fold cross-validation*. Melalui teknik ini, proses pelatihan dan pengujian model dilakukan sebanyak k-kali, sehingga pada setiap iterasi, satu fold berfungsi sebagai set pengujian sementara sisa fold sebagai set pelatihan [30]. Setiap fold akan digunakan secara bergantian sebagai data pelatihan dan validasi untuk melatih dan mengevaluasi model secara menyeluruh, sehingga dapat menghasilkan performa model yang lebih maksimal [31]. Teknik ini juga memastikan distribusi kelas tetap proporsional seperti distribusi dataset asli, sehingga mengurangi risiko bias akibat pembagian data yang tidak merata [32]. Penelitian ini menerapkan nilai k=5 sehingga dataset dibagi menjadi 5-fold dengan 4-fold sebagai data pelatihan dan 1-fold sebagai data pengujian pada setiap iterasi.

2.5. Pelatihan Model

Pada tahap ini, dilakukan pelatihan model pada tugas klasifikasi aspek, klasifikasi sentimen per aspek, dan klasifikasi gabungan aspek-sentimen. Pada tugas klasifikasi aspek, model dilatih untuk mengidentifikasi aspek yang terkandung pada sebuah ulasan secara multilabel. Pada tugas klasifikasi sentimen, model dilatih untuk memprediksi polaritas sentimen terhadap setiap kategori aspek yang teridentifikasi secara terpisah. Sementara itu, pada tugas klasifikasi gabungan aspek-sentimen, model dilatih untuk memprediksi aspek sekaligus sentimennya secara multilabel. Penelitian ini melakukan *hyperparameter tuning* dengan Optuna, yaitu *framework* berbasis *Tree-structured Parzen Estimator*

(TPE) dengan pendekatan *bayesian optimization* [33]. Optuna memungkinkan ruang pencarian hyperparameter yang dinamis sehingga efektif dalam menemukan nilai optimal [34], [35]. Penelitian ini menerapkan pengaturan Optuna n_trials=10, sehingga pencarian hyperparameter pada setiap eksperimen dilakukan sebanyak 10 kali percobaan. Selanjutnya, nilai hyperparameter terbaik yang telah diperoleh digunakan untuk melatih model akhir. Pelatihan model dilakukan menggunakan model indobert-base-p2 sebagai dasar fine-tuning yang tersedia pada platform Hugging Face.

2.5.1. IndoBERT

IndoBERT merupakan model berbasis *Bidirectional Encoder Representations from Transformers* (BERT) yang dikembangkan dan secara khusus dilatih menggunakan korpus besar bahasa Indonesia. Model ini diperkenalkan sebagai bagian dari *Indonesian Natural Language Understanding* (IndoNLU), yang bertujuan untuk menyediakan standar dan sumber daya NLP yang komprehensif bagi bahasa Indonesia. Model IndoBERT dilatih menggunakan dataset Indo4B yang merupakan data teks bahasa Indonesia berskala besar yang mencakup berbagai domain yang diambil dari sosial media, situs web, blog, dan berita [20].

Pada proses *pre-training*, model berbasis BERT menggunakan teknik *Masked Language Model* (MLM) dan *Next Sentence Prediction* (NSP) untuk membangun representasi kontektual [36]. Teknik MLM bertujuan untuk melatih model dalam memahami relasi semantik antar kata dalam kalimat secara dua arah. Sementara itu, NSP berperan untuk membantu model dalam memahami hubungan antar kalimat. Setelah tahap *pre-training*, umumnya dilakukan *fine-tuning* pada IndoBERT. Proses ini tetap mempertahankan struktur dasar BERT, namun menyesuaikan *input* dan *output*-nya dengan kebutuhan tugas tertentu seperti klasifikasi sentimen, *Named Entity Recognition* (NER), dan tugas lainnya. Secara umum, model IndoBERT menunjukkan kinerja yang unggul dalam memahami teks berbahasa Indonesia, terutama dibandingkan dengan model multibahasa seperti mBERT dan XLM-R [20].

2.6. Evaluasi Model

Pada tahap ini, dilakukan evaluasi model IndoBERT untuk melihat seberapa akurat hasil prediksi serta bagaimana kinerja model. Proses evaluasi dilakukan pada ketiga tugas klasifikasi seperti yang dilakukan pada proses pelatihan model, yaitu evaluasi berdasarkan kemampuan model memprediksi aspek, evaluasi kemampuan model dalam mendeteksi sentimen untuk aspek tertentu, dan evaluasi kemampuan model dalam mendeteksi gabungan aspek dan sentimen sekaligus secara benar.

Confusion matrix memuat informasi mengenai hasil klasifikasi berupa data aktual dan data prediksi. Pada tugas klasifikasi sentimen per aspek, evaluasi dilakukan untuk mengetahui efektivitas model pada setiap aspek untuk melakukan klasifikasi secara biner pada dua kelas target, yaitu kelas positif dan negatif. Sementara itu, evaluasi model pada tugas klasifikasi aspek dan klasifikasi gabungan aspek-sentimen dilakukan secara multilabel yaitu dengan *multilabel confusion matrix*. Terdapat empat jenis aspek evaluasi *confusion matrix* pada setiap label C, antara lain *True Positive* (TP_c) atau prediksi positif yang benar, *False Positive* (FP_c) atau prediksi positif yang salah, *True Negative* (TN_c) atau prediksi negatif yang benar dan *False Negative* (FN_c) atau prediksi negatif yang salah [37]. *Confusion matrix* untuk setiap label C diilustrasikan pada Tabel 3.

Tabel 3. *Confusion Matrix* pada Setiap Kelas (C)

	Prediksi Positif	Prediksi Negatif
Aktual Positif	TP_c	FN_c
Aktual Negatif	FP_c	TN_c

Formula *accuracy, precision, recall, dan F1-score* direpresentasikan pada Persamaan (2), (3), (4), dan (5). Nilai *accuracy* dihitung dengan pendekatan *label-based accuracy*, yaitu rata-rata akurasi pada setiap label. Sementara itu, *precision*, *recall*, dan *F1-score* dihitung menggunakan pendekatan *macro averaging* untuk mendapatkan performa keseluruhan model pada klasifikasi multi-label [38].

$$Accuracy = \frac{1}{N}\sum_{c=1}^{N}\frac{TP_c + TN_c}{TP_c + TN_c + FP_c + FN_c} \quad (2)$$

$$Precision_{macro} = \frac{1}{N}\sum_{c=1}^{N}\frac{TP_c}{TP_c + FP_c} \quad (3)$$

$$Recall_{macro} = \frac{1}{N}\sum_{c=1}^{N}\frac{TP_c}{TP_c + FN_c} \quad (4)$$

$$F1_{macro} = \frac{1}{N}\sum_{c=1}^{N}\frac{2 \times Precision_c \times Recall_c}{Precision_c + Recall_c} \quad (5)$$

Pada penelitian ini, kami juga melakukan perbandingan model dengan menguji beberapa model pembanding untuk menguji efektivitas model yang diusulkan. Model klasifikasi serta ekstraksi fitur yang digunakan sebagai pembanding dapat dilihat pada Tabel 4. Seluruh model pembanding dilatih dan dievaluasi ulang menggunakan dataset yang digunakan pada penelitian ini. Model SVM dan XGBoost diimplementasikan menggunakan ekstraksi fitur statis TF-IDF. Sementara itu, model BiLSTM menggunakan *pre-trained* IndoBERT *embeddings*, sedangkan model mBERT dan XLM-R menggunakan ekstraksi fitur yang dihasilkan oleh modelnya masing-masing yang dilakukan *fine-tuning* secara *end-to-end*.

Tabel 4. Model Pembanding dan Metode Ekstraksi Fiturnya

Model	Ekstraksi Fitur
SVM [15], [16]	TF-IDF
XGBoost [9]	TF-IDF
BiLSTM [11]	*Pre-trained* IndoBERT
mBERT [36]	*Fine-tuned* mBERT
XLM-R [39]	*Fine-tuned* XLM-R

2.7. Analisis Hasil

Pada tahap terakhir, hasil dari pengujian model akan dianalisis untuk melihat seberapa efektif model dalam melakukan ketiga tugas klasifikasi. Kemudian, dilakukan analisis kesalahan model berdasarkan *confusion matrix* untuk mengidentifikasi pola tertentu yang menyebabkan model gagal dalam klasifikasi, dengan tujuan untuk memberikan wawasan mengenai keterbatasan model dan potensi pengembangan untuk penelitian selanjutnya.

3. HASIL

3.1. Dataset

Berdasarkan proses pengumpulan data, didapatkan sebanyak 4231 data ulasan pengguna aplikasi Access by KAI. Setelah proses pelabelan, dilakukan pengukuran *interrater agreement* menggunakan *Cohen's Kappa*. Hasil perhitungan *Cohen's Kappa* pada pelabelan aspek dan sentimen ditampilkan pada Tabel 5. Berdasarkan nilai *kappa* antar pelabel, dapat dilihat bahwa kesepakatan label aspek dan sentimen antara pakar dengan penulis dan pakar dengan OpenAI menunjukkan tingkat *substantial* hingga *almost perfect agreement* menurut interpretasi nilai kappa pada Tabel 2. Hal ini mengindikasikan bahwa data telah diberi label secara konsisten dan reliabel. Pada kesepakatan antara penulis dan OpenAI,

Vol. 7, No. 1, February 2026, Page. 286-306
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5402

perhitungan *kappa* pada label sentimen memperoleh nilai sebesar 0.552 dan menunjukkan tingkat *moderate agreement*. Hal ini disebabkan oleh distribusi label yang sangat tidak seimbang, dimana mayoritas data diklasifikasikan ke salah satu kelas saja yaitu sentimen negatif. Sehingga, nilai *kappa* cenderung rendah, meskipun antar pelabel sebenarnya sangat sering sepakat dalam pelabelan yang dibuktikan dengan nilai *percent agreement* sentimen yang menunjukkan nilai sebesar 0.900. Dengan mempertimbangkan kedua metrik ini, dapat disimpulkan bahwa pelabelan aspek dan sentimen pada dataset ini cukup reliabel untuk digunakan dalam tahap permodelan dan analisis lebih lanjut. Dataset pada penelitian ini dapat diakses melalui Github[1].

Tabel 5. Hasil Perhitungan *Interrater Agreement* Ketiga Pelabel

Pasangan *Annotator*	Jumlah Data *Overlap*	Kappa *Agreement* Aspek	*Percent Agreement* Aspek	Kappa *Agreement* Sentimen	*Percent Agreement* Sentimen
Pakar dan Penulis	400	0.914	0.977	0.851	0.990
Pakar dan OpenAI	1000	0.708	0.892	0.781	0.950
Penulis dan OpenAI	600	0.730	0.913	0.552	0.900

Setelah melalui proses pelabelan, selanjutnya dilakukan penghapusan pada data yang tidak memiliki label aspek, serta pada data berlabel sentimen "*Netral*", dan "*Konflik*". Sehingga, didapatkan jumlah data akhir sebanyak 3932 baris data. Sampel data hasil pelabelan ditampilkan pada Tabel 6.

Tabel 6. Sampel Data Hasil Pelabelan

Content	Performa	Tampilan	Tiket	Pembayaran	Akun
Update-nya memang bagus dalam segi tampilan, tapi mobilisasinya jadi melambat. Pembayaran via e-wallet lama loading. Baru-baru ini di halaman untuk booking tiket juga lama, kadang gagal dan harus keluar-masuk apps biar tiketnya bisa 'nyangkut'.	negatif	positif	negatif	negatif	
Aplikasi diupdate bukannya semakin bagus malah semakin jelek, semakin lemot setiap mau liat jadwal kereta KRL ga keluar jadwalnya pdhl udh updatean terbaru. aplikasi jelek 😀😀👎💔	negatif				

3.2. Hasil Pra-pemrosesan Data

Hasil pada setiap proses pra-pemrosesan data meliputi *case folding, cleaning, normalization,* dan *tokenization* ditampilkan pada Tabel 7. Pada tahap terakhir, hasil *tokenization* menghasilkan *token ids* dan *attention masks* yang selanjutnya akan digunakan sebagai input pelatihan model.

Tabel 7. Sampel Hasil Pra-Pemrosesan Data

Proses	Hasil
Teks Awal	Aplikasi diupdate bukannya semakin bagus malah semakin jelek, semakin lemot setiap mau liat jadwal kereta KRL ga keluar jadwalnya pdhl udh updatean terbaru. aplikasi jelek 😀😀👎💔

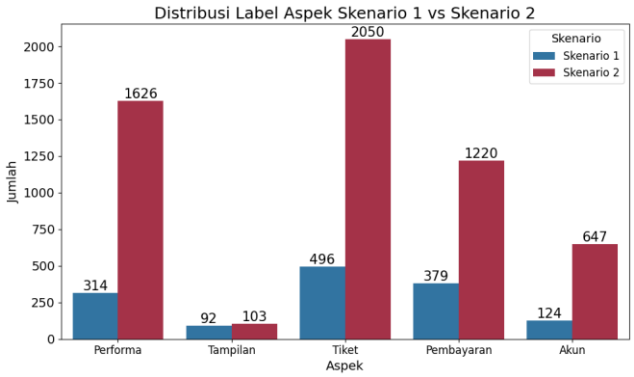
[1] https://github.com/hilldaaa/Dataset-ABSA-AccessbyKAI

Vol. 7, No. 1, February 2026, Page. 286-306
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5402

Case Folding	aplikasi diupdate bukannya semakin bagus malah semakin jelek, semakin lemot setiap mau liat jadwal kereta krl ga keluar jadwalnya pdhl udh updatean terbaru. aplikasi jelek 😀😀👎💔
Cleaning	aplikasi diupdate bukannya semakin bagus malah semakin jelek, semakin lemot setiap mau liat jadwal kereta krl ga keluar jadwalnya pdhl udh updatean terbaru. aplikasi jelek
Normalization	aplikasi diupdate bukannya semakin bagus bahkan semakin jelek, semakin lambat setiap mau lihat jadwal kereta krl tidak keluar jadwalnya padahal sudah updatean terbaru. aplikasi jelek
Tokenization	['[CLS]', 'aplikasi', 'diupdate', 'bukannya', 'semakin', 'bagus', 'bahkan', 'semakin', 'jelek', ',', 'semakin', 'lambat', 'setiap', 'mau', 'lihat', 'jadwal', 'kereta', 'krl', 'tidak', 'keluar', 'jadwal', '##nya', 'padahal', 'sudah', 'update', '##an', 'terbaru', '.', 'aplikasi', 'jelek', '[SEP]', '[PAD]', ...] [2, 1339, 25989, 8003, 977, 1305, 816, 977, 8302, 30468, 977, 5163, 551, 422, 1173, 2864, 3593, 24569, 119, 654, 2864, 57, 2775, 2056, 259, 4398, 5, 1114, 30470, 1339, 8302, 3, 0, ...] [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, ...]

3.3. Klasifikasi Aspek

Pada tahap ini, implementasi model IndoBERT dilakukan pada tugas klasifikasi aspek untuk mengidentifikasi kategori aspek yang muncul dalam setiap ulasan secara multilabel. Penelitian ini melakukan eksperimen klasifikasi aspek pada dua skenario dataset, yaitu data skenario 1 berjumlah 951 baris data yang merupakan hasil pelabelan oleh pakar saja dan data skenario 2 berjumlah 3932 baris data hasil penggabungan pelabelan ketiga pelabel. Distribusi data pada label aspek untuk kedua skenario dataset disajikan pada Gambar 2.



Gambar 2. Distribusi Label Aspek Dataset Skenario 1 dan Skenario 2

Tabel 8. Konfigurasi dan Hasil *Hyperparameter Tuning* Klasifikasi Aspek

Hyperparameter	Nilai Konfigurasi	Nilai Terbaik Data Skenario 1	Nilai Terbaik Data Skenario 2
Learning Rate	1e-6 - 5e-5	3.448e-5	4.970e-6
Dropout	0.1 – 0.5	0.171	0.113
Batch Size	8, 16, 32	16	8

IndoBERT digunakan sebagai *feature extractor* untuk menghasilkan representasi fitur dan ditambahkan lapisan klasifikasi di atasnya. Seluruh *pipeline* model IndoBERT dilakukan optimasi

secara *end-to-end* melalui proses *fine-tuning* menggunakan data dan label spesifik pada tugas klasifikasi aspek. Tabel 8 menampilkan nilai konfigurasi dan hasil *hyperparameter tuning* klasifikasi aspek model IndoBERT. Proses *hyperparameter tuning* IndoBERT dilakukan sebanyak 10 kali percobaan menggunakan *AdamW optimizer* dan *epoch* sebanyak 15 dengan pengaturan *early stopping*.

Tabel 9. Hasil Evaluasi Model Klasifikasi Aspek

Dataset	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Skenario 1	0.894 ± 0.02	0.807 ± 0.02	0.734 ± 0.07	0.758 ± 0.05
Skenario 2	0.928 ± 0.00	0.823 ± 0.05	0.770 ± 0.03	0.785 ± 0.04

Selanjutnya, nilai parameter terbaik digunakan untuk proses pelatihan dan evaluasi untuk mengukur performa model pada kedua skenario data. Hasil performa model IndoBERT pada klasifikasi aspek ditampilkan pada Tabel 9. Berdasarkan hasil metrik evaluasi model, dapat dilihat bahwa dataset skenario 2 secara konsisten menghasilkan nilai metrik yang lebih tinggi dibandingkan dengan dataset skenario 1, dengan perolehan *accuracy* sebesar 0.928 ± 0.00 dan *F1-score* sebesar 0.785 ± 0.04. Peningkatan ini menunjukkan bahwa ukuran data yang lebih besar membantu model untuk lebih menangkap variasi pola dalam ulasan, sehingga performa model dalam memprediksi aspek menjadi semakin baik.

3.3.1. Perbandingan Model Klasifikasi Aspek

Berdasarkan hasil implementasi model yang telah dilakukan, model IndoBERT berhasil memperoleh performa yang baik pada kemampuannya memprediksi kategori aspek. Selanjutnya, kami melakukan perbandingan dengan beberapa model dasar untuk membuktikan efektivitas model IndoBERT. Setiap model pembanding juga telah melalui proses *hyperparameter tuning* dengan Optuna (*n_trials*=10) untuk memastikan perbandingan yang adil. Proses pencarian nilai parameter terbaik dilakukan pada parameter yang relevan untuk masing-masing model.

Tabel 10. Perbandingan Performa Model Klasifikasi Aspek

Dataset	Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Skenario 1	SVM	0.856 ± 0.01	0.776 ± 0.02	0.621 ± 0.03	0.677 ± 0.02
	XGBoost	0.856 ± 0.01	0.756 ± 0.05	0.587 ± 0.02	0.647 ± 0.02
	BiLSTM	0.868 ± 0.01	0.778 ± 0.03	0.653 ± 0.05	0.699 ± 0.04
	mBERT	0.861 ± 0.01	0.711 ± 0.01	0.735 ± 0.07	0.716 ± 0.04
	XLM-R	0.887 ± 0.01	0.772 ± 0.01	0.756 ± 0.09	0.746 ± 0.06
	IndoBERT	0.894 ± 0.02	0.807 ± 0.02	0.734 ± 0.07	0.758 ± 0.05
Skenario 2	SVM	0.907 ± 0.00	0.787 ± 0.02	0.730 ± 0.02	0.757 ± 0.02
	XGBoost	0.912 ± 0.00	0.775 ± 0.09	0.673 ± 0.01	0.700 ± 0.02
	BiLSTM	0.912 ± 0.00	0.821 ± 0.05	0.739 ± 0.03	0.767 ± 0.03
	mBERT	0.921 ± 0.00	0.782 ± 0.05	0.745 ± 0.02	0.748 ± 0.03
	XLM-R	0.925 ± 0.00	0.783 ± 0.07	0.773 ± 0.03	0.768 ± 0.05
	IndoBERT	0.928 ± 0.00	0.823 ± 0.05	0.770 ± 0.03	0.785 ± 0.04

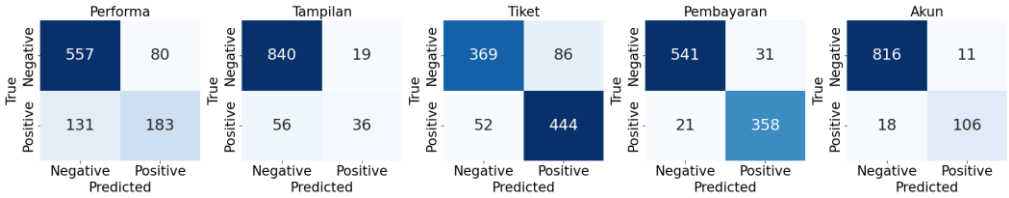
Hasil perbandingan performa model pada klasifikasi aspek ditampilkan pada Tabel 10. Pada data skenario 1, model IndoBERT menunjukkan hasil yang terbaik dan berhasil mengungguli kelima model pembanding. Sementara itu, pada data skenario 2, semua model mengalami peningkatan performa secara keseluruhan. Hal ini disebabkan oleh meningkatnya jumlah sampel data yang digunakan untuk pelatihan, sehingga model dapat lebih memahami pola dan variasi pada data secara keseluruhan. Model IndoBERT secara konsisten tetap menunjukkan hasil yang terbaik. Meskipun nilai recall pada model

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 1, February 2026, Page. 286-306
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5402

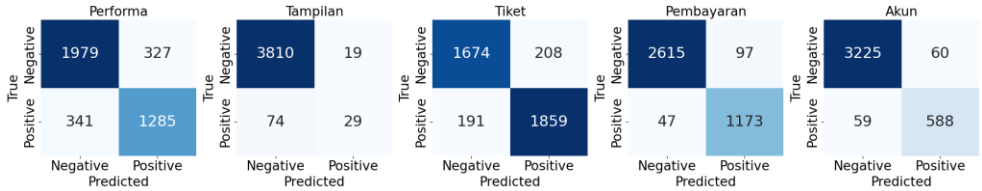
XLM-R sedikit lebih tinggi daripada model IndoBERT, namun model IndoBERT lebih unggul secara keseluruhan karena mampu menyeimbangkan antara precision dan recall sehingga memperoleh nilai F1-score terbaik.

3.3.2. Analisis *Confusion Matrix* Klasifikasi Aspek

Pada tugas klasifikasi aspek, IndoBERT memperoleh performa terbaik dibandingkan semua model pembanding. Selanjutnya, kami melakukan analisis lebih lanjut terhadap model IndoBERT sebagai model terbaik pada tugas ini. Hasil *confusion matrix* model IndoBERT pada tugas klasifikasi aspek disajikan pada Gambar 3 untuk dataset skenario 1 dan Gambar 4 untuk dataset skenario 2.



Gambar 3. *Confusion Matrix* Klasifikasi Aspek Model IndoBERT Dataset Skenario 1

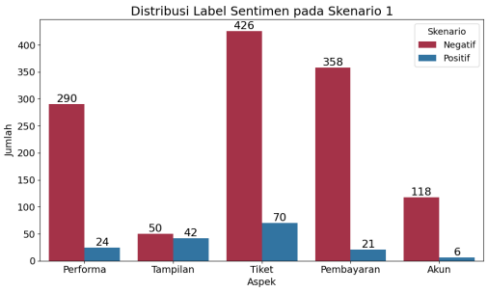


Gambar 4. *Confusion Matrix* Klasifikasi Aspek Model IndoBERT Dataset Skenario 2

Pada kedua skenario dataset, aspek "*Performa*", "*Tiket*", dan "*Pembayaran*" menunjukkan hasil yang paling konsisten dengan jumlah prediksi benar yang cukup tinggi. Sementara itu, pada aspek "*Tampilan*" dan "*Akun*", model masih menunjukkan banyak kesalahan prediksi (*false positive* dan *false negative*) yang mengindikasikan adanya tantangan tertentu dalam membedakan kasus spesifik di kedua aspek tersebut. Hal ini disebabkan oleh distribusi label yang minimum pada kedua aspek tersebut, sehingga proses pembelajaran model menjadi kurang optimal dan risiko kesalahan prediksi menjadi lebih besar.

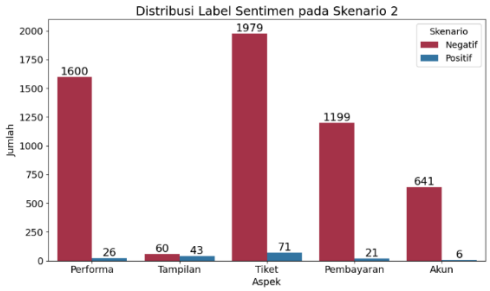
3.4. Klasifikasi Sentimen per Aspek

Pada tahap ini, pendekatan klasifikasi sentimen per aspek dilakukan dengan tujuan mengidentifikasi sentimen positif atau negatif yang terkandung pada setiap aspek secara terpisah. Sama seperti bagian sebelumnya, eksperimen klasifikasi sentimen dilakukan pada dua skenario dataset. Distribusi label sentimen per aspek disajikan pada Gambar 5 untuk dataset skenario 1 dan Gambar 6 untuk dataset skenario 2.



Gambar 5. Distribusi Label Sentimen per Aspek pada Dataset Skenario 1

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 1, February 2026, Page. 286-306
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5402



Gambar 6. Distribusi Label Sentimen per Aspek pada Dataset Skenario 2

Seluruh *pipeline* model IndoBERT melalui proses *fine-tuning* menggunakan data dan label spesifik pada tugas klasifikasi sentimen secara terpisah pada setiap aspek. Sehingga, totalnya terdapat lima model klasifikasi sentimen untuk lima kategori aspek pada setiap skenario dataset. Tabel 11 menampilkan konfigurasi dan hasil *hyperparameter tuning* model klasifikasi sentimen model IndoBERT. Proses Optuna pada IndoBERT dilakukan sebanyak 10 kali percobaan menggunakan *AdamW optimizer* dan *epoch* sebanyak 15 dengan pengaturan *early stopping*.

Tabel 11. *Hyperparameter Tuning* untuk Dataset Skenario 1 (S1) dan Dataset Skenario 2 (S2)

Data	*Hyper-parameter*	Nilai Konfigurasi	Nilai Parameter Terbaik				
			Performa	Tampilan	Tiket	Pembayaran	Akun
S1	*Learning Rate*	1e-6 - 5e-5	1.780e-5	1.894e-5	4.654e-5	1.661e-5	1.377e-6
	Dropout	0.1 – 0.5	0.487	0.329	0.105	0.435	0.117
	Batch Size	8, 16, 32	16	8	16	32	32
S2	*Learning Rate*	1e-6 - 5e-5	1.310e-5	1.072e-5	3.904e-6	1.914e-5	1.444e-5
	Dropout	0.1 – 0.5	0.140	0.373	0.497	0.496	0.268
	Batch Size	8, 16, 32	32	16	16	8	8

Tabel 12. Hasil Evaluasi Model Klasifikasi Sentimen pada Dataset Skenario 1

Dataset	Aspek	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Skenario 1	Performa	0.987 ± 0.01	0.978 ± 0.04	0.938 ± 0.09	0.951 ± 0.05
	Tampilan	0.795 ± 0.09	0.809 ± 0.10	0.786 ± 0.09	0.787 ± 0.10
	Tiket	0.962 ± 0.01	0.949 ± 0.03	0.894 ± 0.05	0.915 ± 0.02
	Pembayaran	0.945 ± 0.01	0.645 ± 0.16	0.608 ± 0.11	0.619 ± 0.13
	Akun	0.952 ± 0.02	0.476 ± 0.01	0.500 ± 0.00	0.488 ± 0.00
	Average	0.928 ± 0.03	0.771 ± 0.07	0.745 ± 0.07	0.752 ± 0.06
Skenario 2	Performa	0.993 ± 0.00	0.964 ± 0.08	0.803 ± 0.08	0.864 ± 0.07
	Tampilan	0.836 ± 0.06	0.840 ± 0.07	0.833 ± 0.06	0.832 ± 0.06
	Tiket	0.991 ± 0.00	0.959 ± 0.03	0.901 ± 0.06	0.925 ± 0.03
	Pembayaran	0.983 ± 0.01	0.643 ± 0.23	0.574 ± 0.11	0.596 ± 0.15
	Akun	0.989 ± 0.00	0.495 ± 0.00	0.499 ± 0.00	0.497 ± 0.00
	Average	0.958 ± 0.02	0.780 ± 0.08	0.722 ± 0.06	0.743 ± 0.06

Selanjutnya, dilakukan proses pelatihan dan evaluasi menggunakan nilai parameter optimal untuk mengukur performa model. Hasil performa model IndoBERT pada klasifikasi sentimen per aspek pada kedua skenario dataset ditampilkan pada Tabel 12. Berdasarkan hasil metrik evaluasi, performa model IndoBERT pada dataset skenario 1 dan 2 menunjukkan karakteristik yang bervariasi pada setiap aspek. Secara keseluruhan, jumlah data yang lebih banyak pada dataset skenario 2 dapat meningkatkan nilai *accuracy* model, namun nilai *F1-score* cenderung menurun dibandingkan dataset skenario 1. Selain

Vol. 7, No. 1, February 2026, Page. 286-306
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5402

itu, model klasifikasi sentimen pada aspek Pembayaran dan Akun menunjukkan performa *F1-score* yang paling rendah dibandingkan dengan aspek lainnya.

3.4.1. Perbandingan Model Klasifikasi Sentimen per Aspek

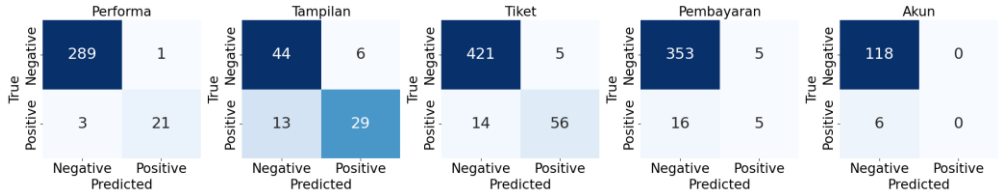
Berdasarkan hasil implementasi model, IndoBERT berhasil memperoleh performa yang cukup baik pada kemampuannya memprediksi kategori sentimen per aspek. Selanjutnya, kami juga melakukan perbandingan dengan beberapa model dasar untuk membuktikan efektivitas model IndoBERT. Seluruh model pembanding telah melalui proses *hyperparameter tuning* dengan Optuna sebanyak 10 kali percobaan, serta dilakukan pelatihan dan evaluasi pada kedua skenario dataset.

Tabel 13. Perbandingan Performa Model Klasifikasi Sentimen per Aspek

Data	Model	*Avg Accuracy*	*Avg Precision*	*Avg Recall*	*Avg F1-Score*
Skenario 1	SVM	0.903 ± 0.03	0.695 ± 0.07	0.634 ± 0.06	0.645 ± 0.06
	XGBoost	0.906 ± 0.03	0.704 ± 0.05	0.662 ± 0.04	0.674 ± 0.04
	BiLSTM	0.922 ± 0.03	0.795 ± 0.08	0.733 ± 0.06	0.746 ± 0.06
	mBERT	0.881 ± 0.04	0.687 ± 0.12	0.647 ± 0.08	0.646 ± 0.10
	XLM-R	0.873 ± 0.03	0.682 ± 0.11	0.674 ± 0.10	0.667 ± 0.10
	IndoBERT	0.928 ± 0.03	0.771 ± 0.07	0.745 ± 0.07	0.752 ± 0.06
Skenario 2	SVM	0.928 ± 0.01	0.714 ± 0.09	0.627 ± 0.06	0.647 ± 0.07
	XGBoost	0.941 ± 0.03	0.720 ± 0.08	0.620 ± 0.05	0.641 ± 0.06
	BiLSTM	0.951 ± 0.02	0.767 ± 0.08	0.715 ± 0.08	0.731 ± 0.07
	mBERT	0.919 ± 0.03	0.717 ± 0.11	0.628 ± 0.07	0.650 ± 0.08
	XLM-R	0.932 ± 0.02	0.704 ± 0.06	0.667 ± 0.06	0.677 ± 0.06
	IndoBERT	0.958 ± 0.02	0.780 ± 0.08	0.722 ± 0.06	0.743 ± 0.06

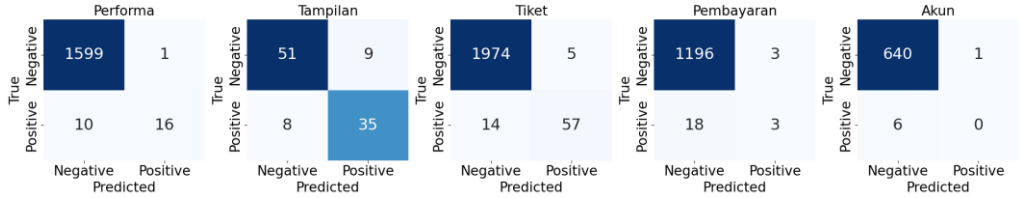
Untuk mempermudah proses perbandingan performa antar model, pada Tabel 13 disajikan nilai rata-rata performa model klasifikasi sentimen untuk kelima aspek. Hasil tersebut menunjukkan bahwa model IndoBERT secara konsisten menjadi model terbaik dan mampu mengungguli model-model pembanding pada kedua skenario data. Jumlah data yang lebih banyak pada dataset skenario 2 secara keseluruhan mampu meningkatkan nilai *accuracy* pada semua model, namun cenderung menurunkan nilai *recall* dan *F1-score.* Hal ini disebabkan oleh ketidakseimbangan distribusi label yang semakin terlihat pada data skenario 2, dimana selisih antara data sentimen positif dan negatif lebih tinggi dibandingkan pada data skenario 1. Kondisi tersebut menyebabkan model cenderung lebih bias terhadap kelas mayoritas dan kesulitan untuk mendeteksi kelas minoritas pada data skenario 2.

3.4.2. Analisis *Confusion Matrix* Klasifikasi Sentimen per Aspek



Gambar 7. *Confusion Matrix* Klasifikasi Sentimen per Aspek IndoBERT Data Skenario 1

Vol. 7, No. 1, February 2026, Page. 286-306
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5402

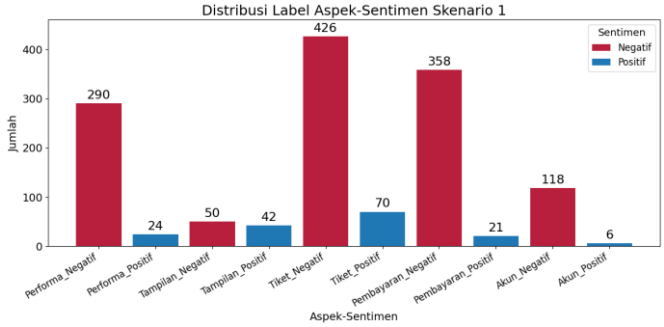


Gambar 8. *Confusion Matrix* Klasifikasi Sentimen per Aspek IndoBERT Data Skenario 2

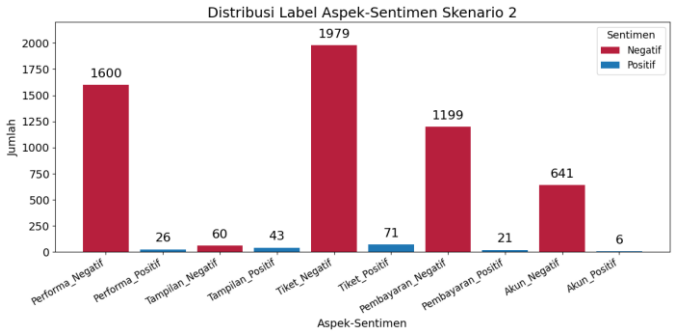
Pada tugas klasifikasi sentimen per aspek, kami juga melakukan analisis terhadap confusion matrix model IndoBERT sebagai model terbaik. Hasil confusion matrix klasifikasi sentimen dengan model IndoBERT disajikan pada Gambar 7 untuk dataset skenario 1 dan Gambar 8 untuk dataset skenario 2.

Secara keseluruhan, klasifikasi sentimen pada aspek "*Performa*", "*Tampilan*", dan "*Tiket*" secara konsisten memperoleh jumlah prediksi benar yang cukup tinggi untuk sentimen negatif maupun positif. Pada aspek "*Pembayaran*", kinerja model sedikit menurun, terlihat dari meningkatnya jumlah *false positive* dan *false negative*. Hal ini menunjukkan bahwa pada aspek ini, model cenderung belum maksimal dalam membedakan sentimen positif dan negatif. Sementara itu pada aspek "*Akun*", sentimen positif sama sekali tidak berhasil diprediksi oleh model IndoBERT di kedua skenario. Hal ini disebabkan oleh jumlah data pelatihan bersentimen positif yang memang sangat sedikit pada aspek tersebut, sehingga model cenderung kesulitan dalam mempelajari pola karakteristik sentimen tersebut.

3.5. Klasifikasi Gabungan Aspek-Sentimen



Gambar 9. Distribusi Label Gabungan Aspek-Sentimen Dataset Skenario 1



Gambar 10. Distribusi Label Gabungan Aspek-Sentimen Dataset Skenario 2

Pada tahap ini, pendekatan klasifikasi gabungan aspek-sentimen dilakukan dengan tujuan mengidentifikasi kategori aspek dan kategori sentimennya sekaligus dalam setiap ulasan secara multilabel. Sama seperti bagian sebelumnya, eksperimen klasifikasi sentimen dilakukan pada dua

skenario dataset. Distribusi data pada label gabungan aspek-sentimen disajikan pada Gambar 9 untuk dataset skenario 1 dan Gambar 10 untuk dataset skenario 2.

Model IndoBERT melalui proses *fine-tuning* menggunakan data dan label spesifik pada tugas klasifikasi gabungan aspek-sentimen. Tabel 14 menampilkan konfigurasi dan hasil *hyperparameter tuning* model klasifikasi gabungan aspek-sentimen model IndoBERT. Proses *hyperparameter tuning* pada IndoBERT dilakukan sebanyak 10 kali percobaan menggunakan *AdamW optimizer* dan *epoch* sebanyak 15 dengan pengaturan *early stopping*.

Tabel 14. Konfigurasi dan Hasil *Hyperparameter Tuning* Klasifikasi Gabungan Aspek-Sentimen

Hyperparameter	Nilai Konfigurasi	Nilai Terbaik Dataset Skenario 1	Nilai Terbaik Dataset Skenario 2
Learning Rate	1e-6 - 5e-5	1.294e-5	1.431e-5
Dropout	0.1 – 0.5	0.363	0.135
Batch Size	8, 16, 32	8	16

Setelah dilakukan *hyperparameter tuning*, nilai parameter optimal digunakan untuk proses pelatihan dan evaluasi untuk mengukur performa model pada kedua skenario dataset. Hasil performa model IndoBERT pada klasifikasi gabungan aspek-sentimen ditampilkan pada Tabel 15. Berdasarkan hasil evaluasi model IndoBERT, dapat dilihat bahwa dataset skenario 2 secara konsisten menghasilkan nilai performa model yang lebih tinggi daripada dataset skenario 1. Hal ini menunjukkan bahwa kemampuan model mengenali dan memprediksi label yang benar menjadi semakin baik pada jumlah dataset yang lebih besar.

Tabel 15. Hasil Evaluasi Model Klasifikasi Gabungan Aspek-Sentimen

Dataset	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Skenario 1	0.935 ± 0.00	0.622 ± 0.08	0.469 ± 0.04	0.500 ± 0.03
Skenario 2	0.962 ± 0.00	0.675 ± 0.08	0.515 ± 0.02	0.549 ± 0.03

3.5.1. Perbandingan Model Klasifikasi Gabungan Aspek-Sentimen

Pada klasifikasi gabungan aspek-sentimen, kami juga melakukan perbandingan dengan beberapa model dasar untuk membuktikan efektivitas model IndoBERT. Seluruh model dasar juga melalui proses *hyperparameter tuning* dengan Optuna sebanyak 10 kali percobaan untuk memastikan perbandingan yang adil.

Hasil perbandingan performa model pada klasifikasi gabungan aspek-sentimen ditampilkan pada Tabel 16. Pada tugas klasifikasi gabungan, model IndoBERT masih menunjukkan performa paling tinggi dibandingkan semua model dasar untuk kedua skenario dataset. Namun, performa seluruh model pada tugas klasifikasi ini pada kedua skenario data masih cukup rendah jika dibandingkan dengan tugas klasifikasi aspek dan tugas klasifikasi sentimen secara terpisah. Hal ini disebabkan oleh meningkatnya kompleksitas tugas klasifikasi ketika aspek dan sentimen harus diprediksi secara bersamaan dengan satu model. Pada kedua skenario, model dasar SVM dan XGBoost secara konsisten menghasilkan performa paling rendah diantara yang lainnya. Model BiLSTM yang menggunakan pre-trained IndoBERT embedding menghasilkan nilai F1-score yang paling baik dibandingkan model pembanding lainnya, tetapi masih tertinggal dari IndoBERT. Sementara itu, mBERT dan XLM-R sebagai model multibahasa berbasis BERT, memiliki nilai accuracy yang paling mendekati model IndoBERT dibandingkan model pembanding lainnya.

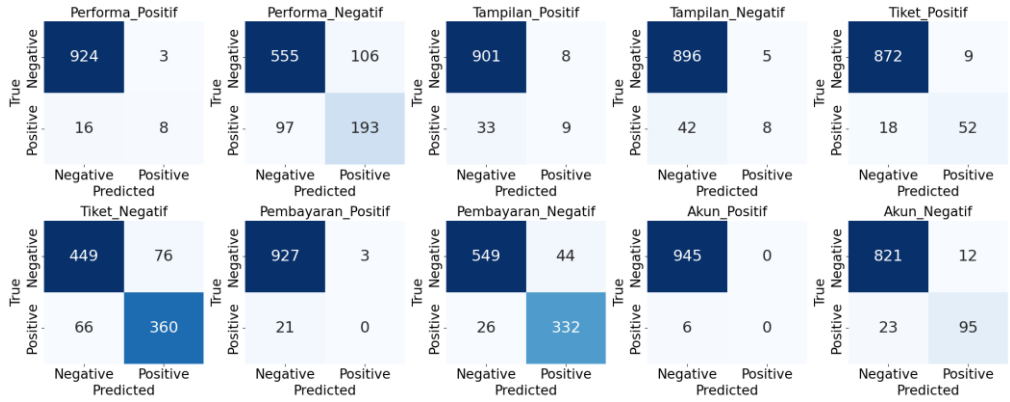
Vol. 7, No. 1, February 2026, Page. 286-306
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5402

Tabel 16. Perbandingan Performa Model Klasifikasi Gabungan Aspek-Sentimen

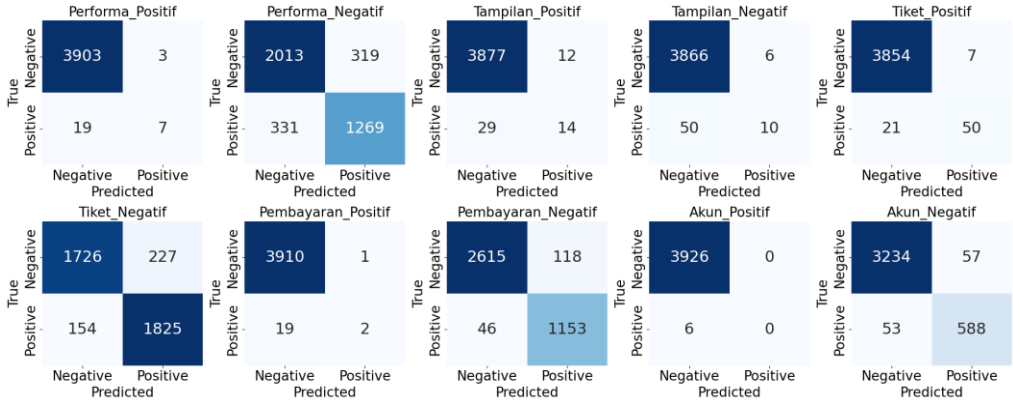
Dataset	Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Skenario 1	SVM	0.918 ± 0.00	0.575 ± 0.06	0.414 ± 0.03	0.455 ± 0.03
	XGBoost	0.912 ± 0.00	0.535 ± 0.05	0.334 ± 0.02	0.391 ± 0.02
	BiLSTM	0.927 ± 0.00	0.549 ± 0.04	0.469 ± 0.03	0.497 ± 0.03
	mBERT	0.929 ± 0.01	0.602 ± 0.06	0.453 ± 0.04	0.483 ± 0.04
	XLM-R	0.935 ± 0.01	0.568 ± 0.09	0.466 ± 0.05	0.483 ± 0.04
	IndoBERT	0.935 ± 0.00	0.622 ± 0.08	0.469 ± 0.04	0.500 ± 0.03
Skenario 2	SVM	0.944 ± 0.00	0.578 ± 0.04	0.410 ± 0.02	0.449 ± 0.02
	XGBoost	0.952 ± 0.00	0.578 ± 0.07	0.405 ± 0.01	0.450 ± 0.02
	BiLSTM	0.951 ± 0.00	0.612 ± 0.04	0.470 ± 0.03	0.504 ± 0.03
	mBERT	0.956 ± 0.00	0.542 ± 0.07	0.459 ± 0.05	0.466 ± 0.03
	XLM-R	0.961 ± 0.00	0.539 ± 0.04	0.465 ± 0.01	0.474 ± 0.01
	IndoBERT	0.962 ± 0.00	0.675 ± 0.08	0.515 ± 0.02	0.549 ± 0.03

3.5.2. Analisis *Confusion Matrix* Klasifikasi Gabungan Aspek-Sentimen

Pada tugas klasifikasi gabungan aspek-sentimen, kami juga melakukan analisis terhadap *confusion matrix* model IndoBERT sebagai model terbaik. Hasil *confusion matrix* klasifikasi gabungan aspek-sentimen pada model IndoBERT disajikan pada Gambar 11 untuk dataset skenario 1 dan Gambar 12 untuk dataset skenario 2.



Gambar 11. *Confusion Matrix* Klasifikasi Gabungan Aspek-Sentimen IndoBERT Data Skenario 1



Gambar 12. *Confusion Matrix* Klasifikasi Gabungan Aspek-Sentimen IndoBERT Data Skenario 2

Pada kedua skenario data, kelas "*Performa_Negatif*", "*Tiket_Negatif*", "*Pembayaran_Negatif*", dan "*Akun_Negatif*" cenderung mendominasi jumlah prediksi benar. Sementara itu, kelas "*Akun_Positif*" secara konsisten pada kedua skenario data, sama sekali tidak memiliki prediksi positif yang benar. Kelas "*Pembayaran_Positif*" juga menunjukkan kesalahan prediksi yang cukup besar. Secara keseluruhan, hasil menunjukkan bahwa model jauh lebih akurat dalam memprediksi gabungan aspek-sentimen yang merupakan kelas mayoritas. Sedangkan pada kelas minoritas, khususnya sentimen positif pada beberapa aspek, menjadi sumber utama kesalahan prediksi.

4. DISKUSI

Berdasarkan hasil implementasi dan perbandingan model yang telah dilakukan, model IndoBERT berhasil memperoleh performa terbaik dibandingkan seluruh model pembanding pada ketiga tugas klasifikasi ABSA terhadap ulasan aplikasi Access by KAI. Performa yang kuat pada tugas klasifikasi aspek mencerminkan kemampuan model dalam memprediksi beberapa kategori aspek dalam sebuah ulasan secara akurat pada skenario multilabel. Pada tugas klasifikasi sentimen per aspek, performa model IndoBERT juga mencerminkan kemampuan yang baik dalam mengidentifikasi sentimen pada setiap aspek yang muncul. Hal serupa juga ditunjukkan pada tugas klasifikasi gabungan aspek-sentimen, di mana model mampu mengidentifikasi kategori aspek sekaligus sentimen yang terkandung dalam sebuah ulasan dengan cukup baik. Secara keseluruhan, hasil penelitian ini membuktikan bahwa performa model *fine-tuning* IndoBERT mampu mengungguli model-model pembanding yaitu SVM, XGBoost, BiLSTM, mBERT, dan XLM-R. Sebagai model yang dilatih khusus dengan data berbahasa Indonesia, IndoBERT mampu memberikan representasi yang lebih kontekstual dan akurat, sehingga dapat menangani kompleksitas bahasa Indonesia pada ulasan aplikasi. Temuan ini juga mengindikasikan bahwa representasi fitur dari model *fine-tuning* IndoBERT lebih baik dibandingkan dengan fitur berbasis TF-IDF, *pre-trained* IndoBERT *embeddings*, serta representasi multibahasa mBERT dan XLM-R. Hasil penelitian ini konsisten dengan penelitian [18] yang menunjukkan bahwa IndoBERT lebih efektif dibandingkan dengan LSTM dan SVM dengan TF-IDF, serta penelitian [19] yang melaporkan keunggulan IndoBERT dibandingkan dengan mBERT dan pendekatan yang menggunakan *pre-trained* IndoBERT *embeddings*. Temuan ini juga sejalan dengan penelitian [21] yang membuktikan bahwa IndoBERT lebih efektif dibandingkan model SVM, LSTM dengan *word embeddings*, dan mBERT.

Meskipun memperoleh performa terbaik pada seluruh tugas, model IndoBERT masih menunjukkan keterbatasan terkait kesalahan klasifikasi. Analisis kesalahan prediksi pada model melalui *confusion matrix* mengungkapkan adanya pola bias yang konsisten pada ketiga tugas klasifikasi, khususnya pada label dengan jumlah data yang terbatas. Model IndoBERT belum sepenuhnya mampu memahami fitur dan pola dari label yang jarang muncul. Hal ini merefleksikan tantangan umum pada ABSA multilabel dalam bahasa Indonesia, di mana distribusi label yang tidak seimbang menyebabkan model cenderung bias terhadap kelas mayoritas. Temuan ini sejalan dengan penelitian [40] yang menunjukkan penurunan kinerja model pada aspek dengan jumlah sampel yang terbatas. Penelitian [41] juga menunjukkan bahwa ketidakseimbangan data menyebabkan model kesulitan dalam memahami karakteristik dari kelas minoritas sehingga meningkatkan risiko kesalahan prediksi. Hal ini mengindikasikan bahwa ketidakseimbangan data masih menjadi faktor yang membatasi kinerja model dalam ABSA bahasa Indonesia, termasuk pada domain ulasan aplikasi dengan distribusi opini yang sangat bervariasi. Strategi penanganan ketidakseimbangan data dapat membantu meningkatkan performa model pada klasifikasi multilabel [42].

Penelitian ini menerapkan pendekatan klasifikasi sentimen dengan mempertimbangkan dua kategori sentimen saja, yaitu positif dan negatif. Beberapa penelitian sebelumnya [15], [16], [17] juga menerapkan skema serupa dengan hanya memfokuskan analisis pada dua label sentimen tersebut. Pada penelitian ini, pengecualian label sentimen netral dan konflik didasarkan pada persentase data yang

sangat kecil, sehingga berpotensi menimbulkan bias pada model. Namun, pembatasan ini menyebabkan hasil yang diperoleh belum sepenuhnya menggambarkan keragaman opini pengguna. Keberadaan label sentimen netral dan konflik memungkinkan pemodelan yang lebih komprehensif terhadap kompleksitas opini pengguna yang muncul dalam kondisi nyata, khususnya pada ulasan yang mengandung ekspresi netral dan ekspresi campuran terhadap aspek tertentu [24]. Oleh karena itu, penelitian di masa mendatang dapat mengeksplorasi strategi dalam mengembangkan pendekatan dengan mempertimbangkan sentimen netral dan konflik.

Temuan pada penelitian ini dapat memberikan implikasi yang signifikan kepada pihak terkait, terutama pengembang dan pengguna aplikasi Access by KAI. Model analisis sentimen berbasis aspek memungkinkan pemahaman yang lebih mendalam mengenai persepsi pengguna pada berbagai aspek layanan dan fitur utama pada aplikasi, seperti performa, tampilan, tiket, pembayaran, dan akun. Wawasan granular ini dapat dimanfaatkan oleh pihak terkait dalam pengambilan keputusan berdasarkan data untuk meningkatkan *user experience* (UX) serta kualitas aplikasi secara keseluruhan. Selain itu, penelitian ini juga berkontribusi pada bidang ilmu komputer, dengan menyediakan dataset ABSA dari ulasan pengguna aplikasi Access by KAI yang dapat diakses secara publik guna mendukung penelitian dan pengembangan lebih lanjut dalam bidang NLP bahasa Indonesia.

5. KESIMPULAN

Penelitian ini bertujuan untuk mengevaluasi kinerja model IndoBERT dalam melakukan tugas klasifikasi aspek, klasifikasi sentimen per aspek, dan klasifikasi gabungan aspek-sentimen pada *Aspect-Based Sentiment Analysis* (ABSA) terhadap ulasan aplikasi Access by KAI. Berdasarkan hasil implementasi dan perbandingan dengan model-model pembanding pada berbagai eksperimen yang telah dilakukan, dapat disimpulkan bahwa IndoBERT merupakan model dengan performa terbaik pada ketiga tugas klasifikasi. Pada tugas klasifikasi aspek, IndoBERT mencapai performa terbaik pada dataset skenario 2 dengan *accuracy* 0.928 dan *F1-score* 0.785. Sementara itu, pada tugas klasifikasi sentimen per aspek, dataset skenario 1 memberikan kinerja terbaik dengan *accuracy* 0.928 dan *F1-score* 0.752. Pada tugas klasifikasi gabungan aspek sentimen, dataset skenario 2 kembali memberikan performa terbaik dengan *accuracy* 0.962 dan *F1-score* 0.549. Secara keseluruhan, model IndoBERT mampu mengungguli model-model pembanding, yaitu Model SVM dan XGBoost dengan ekstraksi fitur TF-IDF, model BiLSTM dengan *pre-trained* IndoBERT *embeddings*, serta model multibahasa yaitu mBERT dan XLM-R. Keunggulan ini menunjukkan efektivitas model IndoBERT sebagai model monolingual yang dilatih spesifik pada data berbahasa Indonesia, sehingga mampu menghasilkan representasi yang lebih akurat. Penelitian ini berkontribusi pada bidang ilmu komputer melalui penerapan analisis sentimen secara granular yang memungkinkan pengembang aplikasi untuk dapat memahami persepsi pengguna secara lebih mendalam pada setiap aspek layanan aplikasi, sehingga dapat dimanfaatkan dalam peningkatan *user experience* (UX). Selain itu, penelitian ini juga menyediakan dataset ABSA *open-source*, yang dapat dimanfaatkan untuk penelitian dan pengembangan lebih lanjut dalam bidang *Natural Language Processing* (NLP). Penelitian selanjutnya dapat melakukan eksplorasi teknik penyeimbangan data seperti *oversampling* atau augmentasi data untuk memperbaiki masalah ketidakseimbangan distribusi label. Selain itu, penelitian selanjutnya juga disarankan untuk melakukan eksplorasi pada model Transformer lainnya seperti IndoBERTweet atau penggabungan beberapa model agar memperoleh hasil yang lebih baik.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

Vol. 7, No. 1, February 2026, Page. 286-306
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5402

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the Institute for Research and Community Service of Universitas Sebelas Maret for supporting this work through the Hibah Penguatan Kapasitas Grup Riset (PKGR-UNS) A funding scheme, Contract Number: 371/UN27.22/PT.01.03/2025. This study is part of the ongoing research initiative of the Data, Information, Knowledge, and Engineering (DIKE) Research Group at Universitas Sebelas Maret.

REFERENCES

[1] Badan Pusat Statistik Indonesia, "Jumlah Penumpang Kereta Api, 2024." Accessed: Mar. 06, 2025. [Online]. Available: https://www.bps.go.id/id/statistics-table/2/NzIjMg==/jumlah-penumpang-kereta-api.html

[2] S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, "Discrepancy detection between actual user reviews and numeric ratings of Google App store using deep learning," *Expert Systems with Applications*, vol. 181, p. 115111, 2021, doi: https://doi.org/10.1016/j.eswa.2021.115111.

[3] M. Demircan, A. Seller, F. Abut, and M. F. Akay, "Developing Turkish sentiment analysis models using machine learning and e-commerce data," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 202–207, June 2021, doi: 10.1016/j.ijcce.2021.11.003.

[4] G. Kontonatsios *et al.*, "FABSA: An aspect-based sentiment analysis dataset of user reviews," *Neurocomputing*, vol. 562, p. 126867, Dec. 2023, doi: 10.1016/j.neucom.2023.126867.

[5] Dyah Ayu Wulandari, Fitra Abdurrachman Bachtiar, and Indriati, "Aspect Based Sentiment Analysis on Shopee Application Reviews Using Support Vector Machine," *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, vol. 15, no. 02, pp. 99–111, Oct. 2025, doi: 10.24843/LKJITI.2024.v15.i02.p03.

[6] M. A. Palimbani, R. P. Hastuti, and R. A. Rajagede, "Analisis Sentimen Berbasis Aspek Pada Ulasan Pengguna Aplikasi Starbucks Menggunakan Algoritma Support Vector Machine," *Journal of Internet and Software Engineering*, vol. 5, no. 1, pp. 43–49, May 2024, doi: 10.22146/jise.v5i1.9130.

[7] N. Yuniar and A. Musdholifah, "Multiclassifier for Aspect-Based Sentiment Analysis on Indonesian Reviews of Kredit Pintar Online Lending App," in *2024 12th International Conference on Information and Communication Technology (ICoICT)*, 2024, pp. 382–389. doi: 10.1109/ICoICT61617.2024.10698093.

[8] M. Ishaq, D. Lestari, and M. Marchenko, "Implementation of Aspect-Based Sentiment Analysis on the Mitra Darat App User Reviews Using Machine Learning," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 4, no. 6, pp. 2763–2776, 2024, doi: 10.30865/klik.v4i6.1889.

[9] Y. Azarya and I. Budi, "Analisis Sentimen Berbasis Aspek Aplikasi Brimo berdasarkan Ulasan Pengguna di Google Playstore," *The Indonesian Journal of Computer Science*, vol. 14, no. 1, pp. 1111–1125, Feb. 2025, doi: 10.33022/ijcs.v14i1.4613.

[10] R. Hardiartama, A. A. Arifiyanti, and S. F. A. Wati3, "Application of Ensemble Machine Learning Methods for Aspect-Based Sentiment Analysis on User Reviews of the Wondr by BNI App," *Jurnal Teknologi dan Open Source*, vol. 8, no. 1, pp. 97–111, June 2025, doi: 10.36378/jtos.v8i1.4297.

[11] E. Subowo, F. A. Artanto, I. Putri, and W. Umaedi, "BLTSM untuk analisis sentimen berbasis aspek pada aplikasi belanja online dengan cicilan," *Jurnal FASILKOM*, vol. 12, no. 2, pp. 132–140, 2022, doi: https://doi.org/10.37859/jf.v12i2.3759.

[12] P. A. Aritonang, M. E. Johan, and I. Prasetiawan, "Aspect-Based Sentiment Analysis on Application Review using CNN (Case Study : Peduli Lindungi Application)," *Ultima Infosys : Jurnal Ilmu Sistem Informasi*, vol. 13, no. 1, pp. 54–61, 2022, doi: https://doi.org/10.31937/si.v13i1.2684.

[13] N. I. Lestari, S. M. Taib, W. Wibowo, I. A. Aziz, and M. R. Habibi, "Aspect-Based Sentiment Analysis for Mobile App Review Using Convolutional Neural Network (CNN) and Word2Vec,"

Vol. 7, No. 1, February 2026, Page. 286-306
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5402

in *2024 IEEE 7th International Conference on Electrical, Electronics and System Engineering (ICEESE)*, 2024, pp. 1–6. doi: 10.1109/ICEESE62315.2024.10828541.

[14] H. Juandri, Hasmawati, and Bunyamin, "Aspect-level Sentiment Analysis on GoPay App Reviews Using Multilayer Perceptron and Word Embeddings," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, vol. 9, no. 4, Nov. 2024, doi: 10.22219/kinetik.v9i4.2041.

[15] H. Mustakim and S. Priyanta, "Aspect-Based Sentiment Analysis of KAI Access Reviews Using NBC and SVM," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 16, no. 2, p. 113, Apr. 2022, doi: 10.22146/ijccs.68903.

[16] S. H. Maghfiroh, D. E. Ratnawati, and P. P. Adikara, "Analisis Sentimen Berbasis Aspek Menggunakan Support Vector Machine dan Binary Relevance Terhadap Aplikasi Access by KAI," Universitas Brawijaya, 2024.

[17] F. A. I. Faradita and Y. Sibaroni, "Multi-Aspect Sentiment Analysis on KAI Access App Reviews (Google Play Store) Using CNN-LSTM Method," in *2025 International Conference on Data Science and Its Applications (ICoDSA)*, Institute of Electrical and Electronics Engineers (IEEE), Sept. 2025, pp. 1106–1111. doi: 10.1109/icodsa67155.2025.11157016.

[18] R. A. Pranata, I. Hidayah, and S. A. I. Alfarozi, "Aspect Based Sentiment Analysis of PLN Customer Complaints Data Using BERT to Improve Services," in *2024 16th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 2024, pp. 258–263. doi: 10.1109/ICITEE62483.2024.10808802.

[19] E. Yulianti and N. K. Nissa, "ABSA of Indonesian customer reviews using IndoBERT: single-sentence and sentence-pair classification approaches," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 5, pp. 3579–3589, Oct. 2024, doi: 10.11591/eei.v13i5.8032.

[20] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Suzhou, China: Association for Computational Linguistics, 2020, pp. 843–857. doi: 10.18653/v1/2020.aacl-main.85.

[21] A. Jazuli, Widowati, and R. Kusumaningrum, "Optimizing Aspect-Based Sentiment Analysis Using BERT for Comprehensive Analysis of Indonesian Student Feedback," *Applied Sciences*, vol. 15, no. 1, p. 172, 2025, doi: 10.3390/app15010172.

[22] R. A. Rahman, V. H. Pranatawijaya, and N. N. K. Sari, "Analisis Sentimen Berbasis Aspek pada Ulasan Aplikasi Gojek," *KONSTELASI: Konvergensi Teknologi dan Sistem Informasi*, vol. 4, no. 1, pp. 70–82, 2024, doi: https://doi.org/10.24002/konstelasi.v4i1.8922.

[23] A. R. Putra and D. E. Ratnawati, "Analisis Sentimen Berbasis Aspek pada Aplikasi Mobile Menggunakan Naïve Bayes berdasarkan Ulasan Pengguna Playstore (Studi Kasus : Jconnect Mobile)," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 12, no. 2, pp. 293–300, Apr. 2025, doi: 10.25126/jtiik.2025127556.

[24] N. Nuryani, R. Munir, A. Purwarianti, and D. Puji Lestari, "BERT-Based Model and LLMs-Generated Synthetic Data for Conflict Sentiment Identification in Aspect-Based Sentiment Analysis," *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 20, p. 004, 2025, doi: 10.28945/5439.

[25] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960, doi: 10.1177/001316446002000104.

[26] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, p. 159, Mar. 1977, doi: 10.2307/2529310.

[27] R. Feldman and J. Sanger, *The text mining handbook : advanced approaches in analyzing unstructured data*. Cambridge University Press, 2007.

[28] L. Owen and I. Putra, "NLP Bahasa Indonesia Resources." Accessed: Feb. 21, 2025. [Online]. Available: https://github.com/louisowen6/NLP_bahasa_resources

[29] A. Nayak, H. Timmapathini, K. Ponnalagu, and V. Gopalan Venkoparao, "Domain adaptation challenges of BERT in tokenization and sub-word representations of Out-of-Vocabulary words," in *Proceedings of the First Workshop on Insights from Negative Results in NLP*, Online: Association for Computational Linguistics, 2020, pp. 1–5. doi: 10.18653/v1/2020.insights-1.1.

Vol. 7, No. 1, February 2026, Page. 286-306
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5402

[30] V. Lumumba, D. Sang, G. Njoka, D. Musyimi, and Kavita, "Comparative Analysis of Cross-Validation Techniques: LOOCV, K-folds Cross-Validation, and Repeated K-folds Cross-Validation in Machine Learning Models," *American Journal of Theoretical and Applied Statistics*, vol. 13, pp. 127–137, Oct. 2024, doi: 10.11648/j.ajtas.20241305.13.

[31] S. Prusty, S. Patnaik, and S. K. Dash, "SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer," *Frontiers in Nanotechnology*, vol. 4, p. 972421, Aug. 2022, doi: 10.3389/fnano.2022.972421.

[32] M. T. R, V. K. V, D. K. V, O. Geman, M. Margala, and M. Guduri, "The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification," *Healthcare Analytics*, vol. 4, p. 100247, 2023, doi: https://doi.org/10.1016/j.health.2023.100247.

[33] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Association for Computing Machinery, 2019, pp. 2623–2631. doi: 10.1145/3292500.3330701.

[34] B. Bischl *et al.*, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 2, p. e1484, Mar. 2023, doi: https://doi.org/10.1002/widm.1484.

[35] A. P. Adhi, K. Umuri, and G. Triyono, "SENTIMENT ANALYSIS AND ENTITY DETECTION ON NEWS HEADLINES TO SUPPORT INVESTMENT DECISIONS," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 6, pp. 1801–1810, Dec. 2024, doi: 10.52436/1.jutif.2024.5.6.3434.

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Association for Computational Linguistics, June 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[37] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-Label Confusion Matrix," *IEEE Access*, vol. 10, pp. 19083–19095, 2022, doi: 10.1109/ACCESS.2022.3151048.

[38] M. C. Hinojosa Lee, J. Braet, and J. Springael, "Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores," *Applied Sciences*, vol. 14, no. 21, p. 9863, Oct. 2024, doi: 10.3390/app14219863.

[39] A. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Association for Computational Linguistics, July 2020, pp. 8440–8451. doi: 10.18653/v1/2020.acl-main.747.

[40] E. F. Aprilia, A. A. Arifiyanti, and N. Sembilu, "Aspect-Based Sentiment Analysis on User Perceptions of OVO using Latent Dirichlet Allocation and Support Vector Machine," *Aviation Electronics, Information Technology, Telecommunications, Electricals, and Controls (AVITEC)*, vol. 7, no. 2, p. 163, June 2025, doi: 10.28989/avitec.v7i2.3035.

[41] M. A. A. O. Putri, I. W. Sumarjaya, and I. G. N. L. Wijayakusuma, "Aspect-Based Sentiment Analysis of Reviews for Pandawa Beach Using Naive Bayes and SVM Methods," *Journal of Applied Informatics and Computing*, vol. 9, no. 2, pp. 305–313, Mar. 2025, doi: 10.30871/jaic.v9i2.9083.

[42] 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," *Aviation Electronics, Information Technology, Telecommunications, Electricals, and Controls (AVITEC)*, vol. 7, no. 2, p. 195, June 2025, doi: 10.28989/avitec.v7i2.3023.