JURNAL ILMIAH SAINS TEKNOLOGI DAN INFORMASI
Volume 4, Nomor 3, Juli 2026
e-ISSN: 2964-3104; p-ISSN: 2964-3090, Hal. 576-591
DOI: https://doi.org/10.59024/jiti.v4i3.2494
Tersedia: https://jurnal.alimspublishing.co.id/index.php/JITI

Analisis Sentimen Berbasis Aspek pada Ulasan Aplikasi Layanan Publik

(Studi Kasus Aplikasi MyPertamina Menggunakan Pendekatan Indobert dan Explainable AI (XAI))

Sri Lestari[1*], Mesra Betty Yel[2], Abror Khanif[3]
[1-3]Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika Jakarta, Indonesia
Email: sri.lestari1203@gmail.com[1], mesrabettyyel@stikomcki.ac.id[2], abrorkhanif3299@gmail.com[3]*
**Penulis Korespondensi: sri.lestari1203@gmail.com*

Abstract. *Application quality assessment on Google Play Store is generally based on numerical ratings, but this approach does not always reflect the actual sentiment of users. The difference between ratings and textual reviews has the potential to cause bias in application quality evaluation, especially for public service applications. This study aims to analyze the sentiment of MyPertamina app users in greater depth using an IndoBERT-based Aspect-Based Sentiment Analysis (ABSA) approach and to evaluate the correspondence between text sentiment and user numerical ratings. The research dataset consists of 9947 user reviews that have undergone preprocessing and aspect separation, with a focus on the account and payment aspects. The IndoBERT model was fine-tuned for sentiment classification and applied in an ensemble scheme to improve prediction stability. In addition, the Explainable Artificial Intelligence (XAI) approach using Integrated Gradients from Captum was used to provide interpretations of the model's prediction results. The results of the study show the dominance of negative sentiment in both main aspects, as well as the discovery of inconsistencies between textual sentiment and numerical ratings in some user reviews. These findings indicate that numerical ratings do not fully represent the actual user experience. Thus, this study contributes to the development of a more transparent and accurate aspect-based sentiment analysis, and offers a more comprehensive evaluation approach for improving the quality of digital public services.*

Keywords: *Application Reviews; Explainable AI; IndoBERT; MyPertamina; Sentiment Analysis*.

Abstrak. Penilaian kualitas aplikasi pada Google Play Store umumnya didasarkan pada rating numerik, namun pendekatan tersebut tidak selalu mencerminkan sentimen pengguna yang sebenarnya. Perbedaan antara rating dan isi ulasan tekstual berpotensi menimbulkan bias dalam evaluasi kualitas aplikasi, khususnya pada aplikasi layanan publik. Penelitian ini bertujuan untuk menganalisis sentimen pengguna aplikasi MyPertamina secara lebih mendalam menggunakan pendekatan *Aspect-Based Sentiment Analysis* (ABSA) berbasis IndoBERT serta mengevaluasi kesesuaian antara sentimen teks dan rating numerik pengguna. Dataset penelitian terdiri dari 9947 ulasan pengguna yang telah melalui tahapan *preprocessing* dan pemisahan aspek, dengan fokus pada aspek *account* dan *payment*. Model IndoBERT dilakukan *fine-tuning* untuk klasifikasi sentimen dan diterapkan dalam skema ensemble guna meningkatkan stabilitas prediksi. Selain itu, pendekatan *Explainable Artificial Intelligence* (XAI) menggunakan *Integrated Gradients* dari *Captum* digunakan untuk memberikan interpretasi terhadap hasil prediksi model. Hasil penelitian menunjukkan dominasi sentimen negatif pada kedua aspek utama, serta ditemukannya ketidaksesuaian antara sentimen tekstual dan rating numerik pada sebagian ulasan pengguna. Temuan ini mengindikasikan bahwa rating numerik belum sepenuhnya merepresentasikan pengalaman pengguna yang sebenarnya. Dengan demikian, penelitian ini memberikan kontribusi dalam pengembangan analisis sentimen berbasis aspek yang lebih transparan dan akurat, serta menawarkan pendekatan evaluasi yang lebih komprehensif bagi peningkatan kualitas layanan publik digital.

Kata kunci: Analisis Sentimen; *Explainable* AI; IndoBERT; MyPertamina; Ulasan Aplikasi.

1. LATAR BELAKANG

Transformasi digital di Indonesia telah membawa perubahan signifikan dalam cara masyarakat mengakses berbagai layanan publik. Aplikasi digital menjadi tulang punggung penyediaan layanan yang cepat, efisien, dan mudah dijangkau. Kehadiran aplikasi seperti MyPertamina menunjukkan upaya pemerintah dalam menyediakan layanan berbasis teknologi untuk meningkatkan kenyamanan masyarakat dalam transaksi dan distribusi subsidi energi.

Naskah Masuk: 22 Mei 2026; Revisi: 22 Juni 2026; Diterima: 23 Juli 2026; Terbit: 31 Juli 2026

Namun, semakin tinggi ketergantungan publik pada aplikasi digital, semakin besar pula tuntutan masyarakat terhadap kualitas, keandalan, dan pengalaman penggunaan aplikasi. Ulasan pengguna di platform Google Play Store kemudian menjadi cerminan langsung persepsi publik terhadap performa aplikasi, sekaligus sumber informasi penting untuk evaluasi dan pengembangan lebih lanjut.

Dalam bidang analisis sentimen, kebutuhan untuk memahami opini pengguna tidak lagi cukup hanya pada level kalimat secara keseluruhan. Pendekatan *Aspect-Based Sentiment Analysis* (ABSA) menjadi penting karena memungkinkan identifikasi sentimen berdasarkan aspek spesifik yang menjadi perhatian pengguna dibandingkan analisis sentimen konvensional (Wafda et al., 2025). Penelitian sebelumnya menunjukkan bahwa pendekatan ABSA yang memanfaatkan model pralatih seperti BERT dan variannya memberikan peningkatan dalam kinerja analisis. Studi pada ulasan aplikasi PLN Mobile menunjukkan bahwa kombinasi antara *word embedding* BERT dan algoritma GRU mampu menghasilkan kinerja analisis yang baik (Selamet et al., 2024). Penelitian lain mengenai aplikasi yang sama juga menunjukkan bahwa penggabungan leksikon sentimen bahasa Indonesia dengan IndoBERT meningkatkan kemampuan model dalam memahami konteks ulasan pengguna (Asri et al., 2025). Temuan serupa diperoleh melalui *fine-tuning* IndoBERT, di mana penyesuaian model pada konteks spesifik memberikan peningkatan performa ABSA (Perwira et al., 2025), dengan tetap memanfaatkan kemampuan model berbasis *transformer* dalam menangani data tekstual dan struktur bahasa yang kompleks (Tunstall et al., 2022). Di sisi lain, kajian mengenai *Explainable Artificial Intelligence* (XAI) menekankan bahwa interpretabilitas model diperlukan agar hasil analisis dapat dipahami dan dievaluasi oleh pengguna maupun pengambil keputusan (Dwivedi et al., 2023).

Meskipun aplikasi layanan publik seperti MyPertamina telah menyediakan mekanisme penilaian berupa *rating* dan ulasan di Google Play Store, permasalahan dalam evaluasi kualitas aplikasi masih sering terjadi. Ulasan pengguna menunjukkan berbagai keluhan yang berulang, seperti kendala login, aplikasi yang tidak responsif, gangguan pada proses transaksi, hingga *bug* pada fitur utama. Namun, penilaian dalam bentuk *rating* bintang sering kali tidak selaras dengan isi ulasan teks yang diberikan pengguna. Dalam beberapa kasus, pengguna dapat memberikan *rating* tinggi meskipun menyampaikan keluhan dalam ulasan, atau sebaliknya memberikan *rating* rendah meskipun isi ulasan tidak sepenuhnya menunjukkan sentimen negatif. Ketidaksesuaian antara *rating* numerik dan opini tekstual menyebabkan *rating* tidak selalu menjadi indikator yang andal dalam merepresentasikan persepsi pengguna terhadap kualitas aplikasi (Aljrees et al., 2024; Ayu et al., 2025). Akibatnya, evaluasi kualitas layanan

digital yang hanya bertumpu pada *rating* berpotensi menghasilkan gambaran yang tidak sepenuhnya mencerminkan permasalahan aktual yang dialami pengguna.

Berdasarkan kondisi tersebut, kesenjangan penelitian tidak hanya terletak pada ketersediaan metode analisis sentimen, tetapi juga pada pendekatan evaluasi aplikasi yang masih terlalu bergantung pada *rating* numerik. Sejumlah penelitian terdahulu telah menerapkan analisis sentimen atau ABSA pada berbagai domain aplikasi, namun hasil analisis tersebut belum banyak dimanfaatkan untuk menjelaskan kontradiksi antara *rating* dan isi ulasan pengguna. Di sisi lain, meskipun model berbasis IndoBERT telah menunjukkan kemampuan dalam memahami konteks bahasa Indonesia (Asri et al., 2025; Perwira et al., 2025), interpretabilitas hasil analisis masih menjadi aspek yang perlu dikembangkan. Model *deep learning* yang kompleks dapat menghasilkan prediksi dengan performa tinggi, tetapi keputusan model tetap sulit dipahami apabila tidak dilengkapi mekanisme interpretasi yang memadai (Dwivedi et al., 2023). Kondisi tersebut menunjukkan perlunya pendekatan yang tidak hanya mampu mengekstraksi sentimen dari ulasan teks, tetapi juga menjelaskan faktor yang mendasari suatu prediksi.

Untuk menjawab permasalahan tersebut, penelitian ini mengusulkan pendekatan analisis sentimen berbasis aspek dengan memanfaatkan IndoBERT sebagai model utama dan *Explainable Artificial Intelligence* (XAI) sebagai mekanisme penjelasan hasil analisis. Pendekatan ini dirancang untuk menjadikan teks ulasan sebagai sumber utama evaluasi, sehingga penilaian kualitas aplikasi tidak hanya bergantung pada *rating* numerik yang tidak selalu selaras dengan opini tekstual pengguna. Dengan mengidentifikasi sentimen berdasarkan aspek-aspek spesifik aplikasi, analisis dapat memberikan gambaran yang lebih rinci mengenai sumber kepuasan dan ketidakpuasan pengguna. Integrasi XAI selanjutnya memungkinkan interpretasi terhadap prediksi model sehingga hasil analisis tidak hanya dinilai berdasarkan performa klasifikasi, tetapi juga dapat ditelusuri berdasarkan kontribusi fitur atau token terhadap keputusan model (Dwivedi et al., 2023). Pendekatan ini sejalan dengan temuan mengenai efektivitas IndoBERT untuk ABSA berbahasa Indonesia dan kebutuhan terhadap interpretabilitas dalam penggunaan model kecerdasan buatan (Perwira et al., 2025).

Urgensi penelitian ini terletak pada potensinya dalam memberikan informasi berbasis data kepada pengembang aplikasi dan pemangku kebijakan. Identifikasi aspek yang paling banyak memperoleh sentimen negatif dapat membantu menentukan area layanan yang membutuhkan perbaikan. Pendekatan yang memadukan IndoBERT dan XAI diharapkan tidak hanya menghasilkan analisis sentimen dengan performa yang memadai, tetapi juga memberikan transparansi terhadap dasar prediksi model. Dengan demikian, hasil analisis dapat

dimanfaatkan sebagai salah satu sumber informasi dalam evaluasi dan peningkatan kualitas layanan publik digital.

2. KAJIAN TEORITIS

Aspect-Based Sentiment Analysis

Analisis sentimen merupakan pendekatan dalam *Natural Language Processing* (NLP) untuk mengidentifikasi opini, evaluasi, atau sikap yang diekspresikan dalam teks, yang umumnya diklasifikasikan menjadi sentimen positif, negatif, atau netral. Analisis pada tingkat dokumen atau kalimat memiliki keterbatasan ketika satu ulasan mengandung opini berbeda terhadap beberapa atribut. *Aspect-Based Sentiment Analysis* (ABSA) mengatasi keterbatasan tersebut dengan mengidentifikasi aspek tertentu dan menentukan polaritas sentimen untuk masing-masing aspek, sehingga mampu memberikan informasi yang lebih rinci mengenai sumber kepuasan atau ketidakpuasan pengguna (Liu, 2016; Wafda et al., 2025).

Efektivitas ABSA pada teks berbahasa Indonesia telah ditunjukkan dalam berbagai domain. Wafda et al. (2025) melaporkan akurasi 97% untuk identifikasi aspek dan 85% untuk klasifikasi sentimen menggunakan IndoBERT. Jazuli et al. (2023) memperoleh akurasi 0,890 untuk ekstraksi aspek dan 0,879 untuk analisis sentimen, sedangkan penelitian berikutnya melaporkan peningkatan hingga akurasi ekstraksi aspek 0,973 dan *F1-score* sentimen 0,974 setelah optimasi model BERT pada ulasan berbahasa Indonesia (Jazuli et al., 2025). Temuan tersebut menunjukkan bahwa pendekatan berbasis *transformer* mampu menangkap hubungan kontekstual antara aspek dan ekspresi sentimen lebih baik dibandingkan analisis sentimen yang hanya menghasilkan polaritas umum.

Transformer dan IndoBERT

Arsitektur *transformer* menggunakan mekanisme *self-attention* untuk membangun representasi kontekstual suatu token berdasarkan hubungannya dengan token lain dalam teks. Arsitektur ini menjadi dasar model bahasa pralatih seperti BERT yang selanjutnya dapat disesuaikan melalui proses *fine-tuning* untuk berbagai tugas NLP, termasuk klasifikasi dan analisis sentimen (Tunstall et al., 2022). IndoBERT mengadaptasi pendekatan tersebut untuk korpus berbahasa Indonesia sehingga lebih sesuai dalam menangani struktur, kosakata, dan variasi linguistik bahasa Indonesia.

Penerapan IndoBERT pada ABSA menunjukkan performa yang konsisten pada berbagai jenis ulasan. Yulianti dan Nissa (2024) menunjukkan bahwa pendekatan *sentence-pair* berbasis IndoBERT memberikan peningkatan *F1-score* hingga 23,6% dibandingkan *baseline*, sementara *fine-tuning single-sentence* memberikan keseimbangan antara performa dan

efisiensi komputasi. *Domain-specific fine-tuning* juga memungkinkan model beradaptasi terhadap karakteristik linguistik pada konten pengguna di domain tertentu (Perwira et al., 2025). Pada ulasan aplikasi layanan digital, Fiarni dan Cellose (2024) melaporkan akurasi hingga 95% dan *F1-score* 94%, sedangkan Asri et al. (2025) memperoleh akurasi 81% pada ulasan PLN Mobile setelah menggabungkan leksikon bahasa Indonesia dan IndoBERT. Literatur tersebut mendukung penggunaan IndoBERT sebagai model utama untuk analisis sentimen berbasis aspek pada ulasan aplikasi berbahasa Indonesia.

Ulasan Pengguna dan Ketidaksesuaian *Rating*

Ulasan pengguna merupakan sumber data tekstual yang menggambarkan pengalaman langsung terhadap suatu aplikasi atau layanan. Berbeda dengan *rating* bintang yang hanya memberikan representasi numerik, teks ulasan dapat memuat informasi mengenai fitur, gangguan teknis, pengalaman transaksi, maupun aspek tertentu yang dianggap memuaskan atau bermasalah. Analisis terhadap ulasan tersebut menjadi relevan bagi aplikasi layanan publik karena dapat digunakan untuk mengidentifikasi permasalahan operasional dan kebutuhan perbaikan layanan secara lebih spesifik.

Penggunaan *rating* sebagai indikator tunggal memiliki keterbatasan karena polaritas teks tidak selalu konsisten dengan jumlah bintang yang diberikan. Aljrees et al. (2024) menemukan ketidaksesuaian antara isi ulasan dan *rating* pada sekitar 25,64% data aplikasi. Fenomena serupa ditemukan pada ulasan PLN Mobile, dengan ketidaksesuaian masing-masing 19% pada kelas sentimen positif dan 13% pada kelas negatif (Asri et al., 2025). Temuan tersebut menunjukkan bahwa analisis berbasis teks dapat memberikan informasi tambahan yang tidak sepenuhnya tertangkap melalui *rating* numerik. Dengan demikian, perbandingan antara sentimen tekstual dan *rating* menjadi komponen penting untuk mengevaluasi sejauh mana penilaian numerik merepresentasikan pengalaman pengguna.

Black Box* dan *Explainable Artificial Intelligence

Model berbasis *deep learning* dan *transformer* memiliki struktur parameter dan hubungan internal yang kompleks sehingga sering dipandang sebagai sistem *black box*. Tingginya performa prediktif tidak secara otomatis menjelaskan faktor yang menyebabkan suatu data diklasifikasikan ke dalam kelas tertentu. Keterbatasan tersebut menjadi semakin penting ketika hasil model digunakan untuk mendukung evaluasi layanan atau pengambilan keputusan yang membutuhkan transparansi dan akuntabilitas (Dwivedi et al., 2023).

Explainable Artificial Intelligence (XAI) dikembangkan untuk meningkatkan interpretabilitas keputusan model melalui identifikasi fitur yang memengaruhi suatu prediksi. XAI dapat menghasilkan penjelasan global maupun lokal dan membantu mendeteksi pola, bias,

atau kesalahan yang tidak terlihat hanya dari metrik performa model (Dwivedi et al., 2023). Pada model berbasis bahasa, pendekatan atribusi dapat digunakan untuk mengidentifikasi kata atau token yang berkontribusi terhadap kelas prediksi tertentu. Captum menyediakan berbagai metode atribusi, termasuk *Integrated Gradients*, yang memungkinkan kontribusi token terhadap keluaran model dianalisis dan divisualisasikan (Miglani et al., 2023). Dalam skripsi, metode tersebut digunakan untuk menelusuri kontribusi token terhadap prediksi sentimen sampai tingkat kata.

3. METODE PENELITIAN

Penelitian ini menggunakan dataset ulasan pengguna aplikasi MyPertamina yang dikumpulkan secara mandiri dari Google Play Store melalui *web crawling*, dengan setiap entri memuat teks ulasan, *rating* bintang, aspek, dan label sentimen. Analisis difokuskan pada aspek *account* dan *payment*. Data terlebih dahulu melalui pelabelan sentimen otomatis ke dalam tiga kelas, yaitu positif, netral, dan negatif, kemudian diproses melalui tahapan *filtering* untuk menghapus data duplikat, spam, terlalu pendek, atau tidak relevan, *data cleaning* untuk menghilangkan URL, emoji, simbol, angka, dan tanda baca berlebih, normalisasi kata tidak baku, *case folding*, *stopword removal*, dan *stemming*. Dataset selanjutnya dibagi menjadi data latih, validasi, dan uji dengan rasio 70:10:20. Teks hasil *preprocessing* ditokenisasi menggunakan *tokenizer* IndoBERT berbasis *WordPiece*, kemudian model IndoBERT di-*fine-tune* untuk melakukan *Aspect-Based Sentiment Analysis* (ABSA) dengan memanfaatkan representasi [CLS], lapisan *fully connected*, dan fungsi *softmax*. Kinerja model dievaluasi pada data uji menggunakan *confusion matrix*, *accuracy*, *precision*, *recall*, dan *F1-score* untuk menilai kemampuan klasifikasi sentimen positif, netral, dan negatif pada aspek *account* dan *payment*. Selanjutnya, prediksi sentimen dibandingkan dengan polaritas yang direpresentasikan oleh *rating* Google Play Store melalui *Authenticity Score* untuk mengidentifikasi kesesuaian (*match*) maupun ketidaksesuaian (*drift*) antara sentimen tekstual dan *rating* numerik. Interpretabilitas model dianalisis menggunakan pendekatan *Explainable Artificial Intelligence* (XAI) melalui metode *Integrated Gradients* pada pustaka Captum, yang mengukur kontribusi setiap token terhadap prediksi model berdasarkan perubahan keluaran sepanjang jalur interpolasi dari *baseline* ke input aktual. Evaluasi XAI dilakukan secara kualitatif melalui visualisasi token yang paling berpengaruh terhadap prediksi, sehingga selain menilai performa klasifikasi dan kesesuaian sentimen dengan *rating*, penelitian ini juga mengevaluasi transparansi dasar linguistik yang digunakan model dalam menghasilkan keputusan.

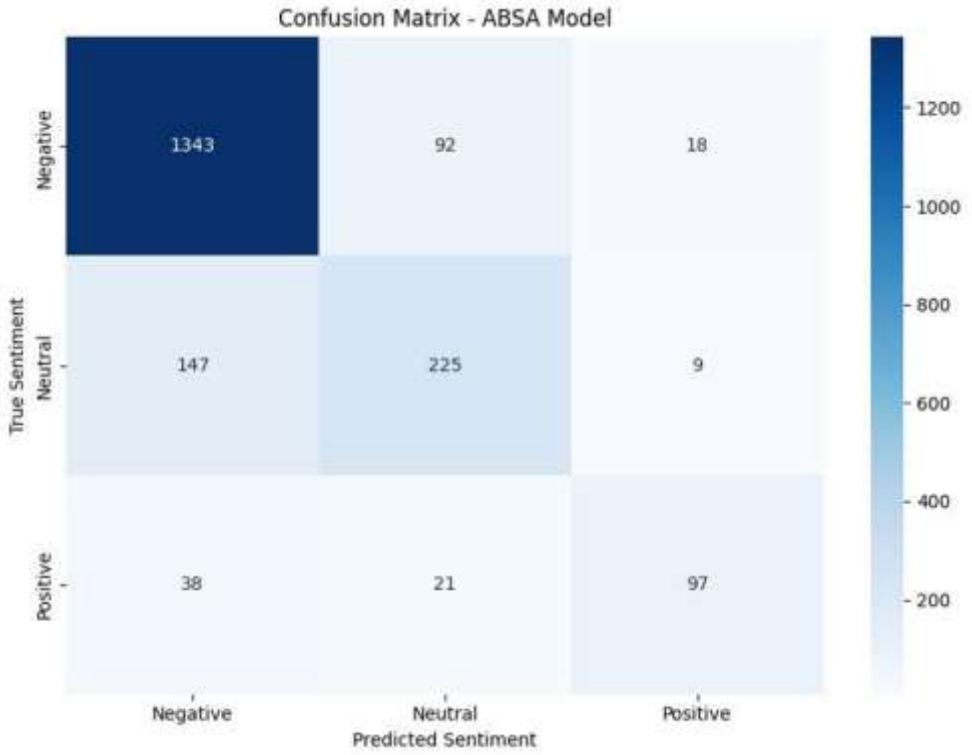
4. HASIL DAN PEMBAHASAN

Pada tahap ini dilakukan pengujian model terhadap data uji yang tidak terlibat dalam proses pelatihan maupun validasi. Tujuan pengujian ini adalah untuk mengevaluasi sejauh mana model mampu melakukan generalisasi dalam mengklasifikasikan sentimen berbasis aspek pada ulasan aplikasi MyPertamina. Evaluasi tidak hanya difokuskan pada nilai metrik kuantitatif, tetapi juga pada distribusi prediksi, konsistensi antar kelas sentimen, serta pola kesalahan yang muncul selama proses klasifikasi.

Melalui tahap ini, dapat diamati bagaimana model merespons variasi bahasa pengguna yang lebih beragam pada data uji, termasuk penggunaan bahasa informal, singkatan, maupun ekspresi emosional yang sering muncul dalam ulasan aplikasi. Hasil pengujian ini kemudian dianalisis untuk memahami kekuatan dan keterbatasan model dalam merepresentasikan persepsi pengguna secara kontekstual, sehingga dapat menjawab rumusan masalah terkait performa model pada tugas *Aspect-Based Sentiment Analysis*.

Evaluasi Model

Untuk mengevaluasi kinerja model secara lebih komprehensif, dilakukan analisis menggunakan *confusion matrix*. Visualisasi ini digunakan untuk melihat distribusi prediksi model terhadap masing-masing kelas sentimen serta mengidentifikasi pola ketepatan dan kesalahan klasifikasi yang terjadi pada data uji.



Gambar 1. *Confusion Matrix.*

Pada Gambar 1 menampilkan *confusion matrix* dari model *Aspect-Based Sentiment Analysis* (ABSA) yang digunakan untuk mengevaluasi kinerja klasifikasi pada tiga kelas sentimen: negatif, netral, dan positif. *Confusion matrix* memperlihatkan perbandingan antara label sebenarnya (*true sentiment*) pada sumbu vertikal dan label hasil prediksi model pada sumbu horizontal, sehingga dapat dianalisis pola ketepatan maupun kesalahan prediksi secara rinci.

Secara visual, intensitas warna pada matriks merepresentasikan jumlah data pada setiap sel. Warna biru yang semakin gelap menunjukkan jumlah prediksi yang semakin besar, sedangkan warna yang lebih terang menunjukkan jumlah yang lebih kecil. Dengan demikian, sel berwarna biru tua menandakan frekuensi kejadian yang tinggi, sementara sel dengan warna pucat menandakan frekuensi rendah. Intensitas warna ini membantu mengidentifikasi secara cepat area dengan konsentrasi prediksi tertinggi maupun pola kesalahan yang dominan.

Pada kelas negatif, terlihat sel diagonal dengan warna biru paling gelap bernilai 1343, yang menunjukkan bahwa sebagian besar data berlabel negatif berhasil diprediksi dengan benar sebagai negatif. Warna yang jauh lebih terang pada sel 92 (diprediksi netral) dan 18 (diprediksi positif) menunjukkan bahwa kesalahan pada kelas ini relatif kecil dibandingkan jumlah prediksi yang benar. Pola ini mengindikasikan bahwa model memiliki kemampuan yang sangat baik dalam mengenali sentimen negatif, kemungkinan karena ekspresi negatif dalam ulasan cenderung eksplisit dan memiliki kata-kata bermuatan emosional kuat.

Pada kelas netral, warna pada sel diagonal (225) tampak lebih terang dibandingkan diagonal kelas negatif, yang menandakan jumlah prediksi benar lebih sedikit. Selain itu, sel dengan nilai 147 pada kolom negatif menunjukkan warna yang cukup pekat, mengindikasikan bahwa cukup banyak data netral yang salah diklasifikasikan sebagai negatif. Pola ini menunjukkan adanya kecenderungan model untuk menginterpretasikan sebagian ekspresi netral sebagai sentimen negatif. Hal ini dapat terjadi karena kalimat netral sering mengandung kritik ringan atau deskripsi masalah tanpa ekspresi emosional yang jelas, sehingga model lebih mudah mengasosiasikannya dengan polaritas negatif.

Pada kelas positif, sel diagonal bernilai 97 menunjukkan bahwa sebagian data positif berhasil diprediksi dengan benar, namun intensitas warnanya masih lebih rendah dibandingkan kelas negatif. Terdapat pula 38 data positif yang diprediksi sebagai negatif dan 21 sebagai netral. Warna pada sel-sel ini menunjukkan bahwa kesalahan prediksi positif ke negatif masih cukup signifikan. Hal ini mengindikasikan bahwa model terkadang gagal membedakan antara ekspresi positif yang lemah dengan netral, atau keliru menangkap konteks sehingga bergeser ke arah negatif.

Secara keseluruhan, pola warna pada *confusion matrix* memperlihatkan dominasi warna gelap pada diagonal kelas negatif, yang menunjukkan kekuatan model dalam mengenali sentimen negatif. Sebaliknya, diagonal kelas netral dan positif memiliki intensitas warna yang lebih moderat, sementara beberapa sel di luar diagonal, khususnya pada pergeseran netral ke negatif, menunjukkan warna yang cukup pekat. Hal ini menandakan bahwa tantangan utama model terletak pada pemisahan kelas netral dan positif dari kelas negatif.

Dengan demikian, *confusion matrix* ini tidak hanya menggambarkan tingkat ketepatan klasifikasi, tetapi juga mengungkap kecenderungan bias prediksi model terhadap kelas tertentu. Analisis warna dan distribusi nilai pada matriks memberikan wawasan yang lebih komprehensif mengenai karakteristik performa model serta area yang masih memerlukan perbaikan pada penelitian lanjutan.

Terdapat hasil evaluasi model berdasarkan sentimen pada tabel 1. berikut:

Tabel 1. Hasil Evaluasi Model (Sentimen).

Sentiment	Precision	Recall	FI-Score	Support
Negative	0,88	0,92	0,90	1453
Neutral	0,67	0,59	0,63	381
Positive	0,78	0,62	0,69	156

Berdasarkan Tabel 1, model menunjukkan performa paling kuat pada kelas negatif, dengan *precision* sebesar 0,88 dan *recall* sebesar 0,92. Nilai *recall* yang lebih tinggi dibanding precision menunjukkan bahwa sebagian besar ulasan negatif berhasil teridentifikasi dengan benar oleh model. Hal ini mengindikasikan bahwa pola keluhan pengguna cenderung memiliki ciri linguistik yang lebih eksplisit, seperti penggunaan kata-kata bernada emosional atau frasa yang secara semantik kuat mengindikasikan ketidakpuasan. Dengan demikian, model relatif sensitif terhadap ekspresi negatif.

Sebaliknya, pada kelas positif, *precision* mencapai 0,78 namun *recall* hanya 0,62. Perbedaan ini menunjukkan bahwa ketika model memprediksi suatu ulasan sebagai positif, prediksi tersebut cukup akurat, tetapi model tidak selalu berhasil menangkap seluruh ulasan positif yang ada. Secara linguistik, hal ini dapat dipengaruhi oleh karakteristik ulasan positif yang lebih singkat, kurang deskriptif, atau menggunakan ekspresi netral yang ambigu, sehingga batas antara sentimen positif dan netral menjadi kurang tegas.

Untuk kelas netral, nilai *precision* dan *recall* yang relatif lebih rendah (0,67 dan 0,59) mengindikasikan bahwa kelas ini merupakan kategori yang paling menantang. Secara konseptual, sentimen netral memang memiliki batas semantik yang lebih kabur dibandingkan negatif dan positif, karena sering kali berupa pernyataan informatif tanpa ekspresi emosional

yang kuat. Kondisi ini menyebabkan model lebih sulit membedakan antara netral dan positif ringan maupun negatif ringan.

Temuan ini menunjukkan bahwa performa model tidak merata di semua kelas, tetapi dipengaruhi oleh karakteristik distribusi data dan pola bahasa pada masing-masing sentimen. Hal ini menunjukkan bahwa performa IndoBERT dalam tugas ABSA cukup kuat pada sentimen negatif, namun masih memiliki ruang perbaikan pada sentimen netral dan positif.

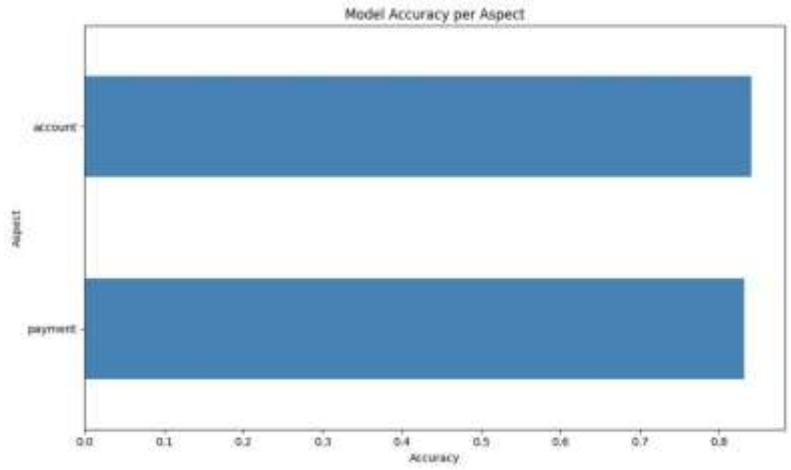
Tabel 2. Hasil Evaluasi Model (Aspek).

Aspect	Total Samples	Accuracy	FI-Score
account	1032	0,8411	0,8259
payment	958	0,8319	0,8323

Berdasarkan Tabel 2, performa model pada aspek *account* dan *payment* berada pada rentang yang relatif berdekatan, dengan *accuracy* masing-masing sebesar 0,8411 dan 0,8319. Selisih yang kecil ini menunjukkan bahwa model tidak menunjukkan kecenderungan bias yang signifikan terhadap salah satu aspek.

Meskipun demikian, jika dikaitkan dengan karakteristik ulasan, aspek *account* umumnya berkaitan dengan isu login, verifikasi, atau registrasi, yang sering disertai keluhan langsung dan ekspresi emosional yang kuat. Hal ini memungkinkan model lebih mudah mengenali polaritas sentimen. Sementara itu, aspek *payment* dapat mengandung variasi konteks yang lebih luas, seperti transaksi berhasil, gagal, atau kendala teknis, sehingga kompleksitas semantik yang lebih tinggi dapat memengaruhi stabilitas prediksi.

Visualisasi pada Gambar 2 di bawah ini memperlihatkan bahwa distribusi akurasi antar aspek relatif seimbang, mengindikasikan bahwa pendekatan ABSA yang digunakan mampu menangani lebih dari satu topik domain tanpa penurunan performa yang signifikan. Dengan demikian, model dapat dikatakan memiliki kemampuan generalisasi yang konsisten pada dua aspek utama yang dianalisis.



Gambar 2. Hasil Akurasi Model (Aspek).

Tabel 3.berikut menunjukkan hasil evaluasi model secara keseluruhan, berbeda dengan evaluasi sebelumnya yang hanya memperlihatkan masing dari hasil evaluasi aspek dan sentimen.

Tabel 3. Hasil Evaluasi Model (Keseluruhan).

Accuracy	0.8367
Precision	0.8305
Recall	0.8367
FI-Score	0.8320

Secara agregat yang terlihat pada Tabel 3, model memperoleh *accuracy* sebesar 0,8367 dengan F1-score sebesar 0,8320. Nilai *precision* dan *recall* yang berada pada kisaran yang sama menunjukkan bahwa model tidak terlalu condong pada kesalahan tipe tertentu (*false positive* atau *false negative*), melainkan mempertahankan keseimbangan antara keduanya.

Namun, apabila dikaitkan dengan distribusi kelas yang tidak sepenuhnya seimbang di mana sentimen negatif mendominasi jumlah sampel maka performa keseluruhan ini juga dipengaruhi oleh kekuatan model dalam mengenali kelas negatif. Oleh karena itu, interpretasi performa tidak hanya bergantung pada nilai *accuracy*, tetapi juga perlu mempertimbangkan variasi performa antar kelas sebagaimana telah dianalisis sebelumnya.

Secara keseluruhan, hasil evaluasi menunjukkan bahwa IndoBERT memiliki kapasitas representasi yang memadai dalam melakukan *Aspect-Based Sentiment Analysis* pada ulasan MyPertamina, dengan performa yang stabil antar aspek namun masih menghadapi tantangan pada sentimen netral dan positif.

Perbandingan Sentimen Dengan *Rating* (*Authenticity Score*)

Perbandingan antara hasil prediksi sentimen berbasis aspek dengan *rating* bintang dilakukan untuk mengevaluasi tingkat konsistensi antara evaluasi numerik dan ekspresi tekstual pengguna. Hasil perhitungan menunjukkan bahwa tingkat kesesuaian (*match rate*) mencapai 0,7678 atau sekitar 76,78%, dengan total 462 kasus *mismatch* dari keseluruhan data uji sebanyak 1990. Nilai ini mengindikasikan bahwa pada sebagian besar ulasan, *rating* bintang dan sentimen tekstual berada pada arah polaritas yang sama, namun masih terdapat proporsi yang signifikan di mana keduanya tidak selaras. Detailnya dapat dilihat pada Gambar 3..

```
Menghitung Authenticity Score...

============================================================
AUTHENTICITY SCORE
============================================================
Match Rate (Predicted vs Rating): 0.7678
============================================================

Total mismatches: 462

Contoh mismatches (5 teratas):
                                                 review   aspect  rating  \
1456                                      Dk bisa login  account       5
7550  Mantap kali MyPertamina, saya mau isi bensin b...  payment       1
1126    Kenapa sekarang menunggu verifikasi lama sekali  account       5
8618  Kenapa harus pake aplikasi kan bayar langsung ...  payment       1
5416  USUL seharusnya aplikasi sperti shopie atau gu...  payment       2

     rating_sentiment pred_label
1456         positive    neutral
7550         negative   positive
1126         positive   negative
8618         negative   positive
5416         negative    neutral
```

Gambar 3. Hasil *Authenticity Score.*

Pada Tabel 4. terdapat sample dari ulasan yang termasuk dari nilai *match* antara hasil analisa dan *rating*.

Tabel 4. Sampel *Authenticity score (Match).*

Review	Sentiment	Rating
kemudahan transaksi, lanjutkan dan kembangkan.	positive	positive
Ngribeti urip. Menghabiskan anggaran untuk uji coba. Rakyat bayar pajak cm dipake untuk uji coba yg meribetkan	negative	negative
Sistem pembayaran tambahkan dong no cash, seperti dana, gopay, dll	neutral	neutral

Jika dilihat pada tabel tersebut, kesesuaian terjadi ketika ekspresi linguistik dalam teks bersifat eksplisit dan tidak ambigu. Ulasan seperti “kemudahan transaksi, lanjutkan dan kembangkan” secara jelas mengandung apresiasi, sehingga selaras dengan rating positif. Demikian pula, ulasan bernada kritik langsung seperti “Ngribeti urip… meribetkan” konsisten dengan *rating* negatif. Pola ini menunjukkan bahwa ketika pengguna mengekspresikan opini secara lugas dan emosional, baik sistem *rating* maupun model ABSA mampu merepresentasikan persepsi tersebut secara konsisten.

Sebaliknya, pada 462 kasus *mismatch* Tabel 5, ditemukan beberapa pola yang mencerminkan kompleksitas interpretasi persepsi pengguna.

Tabel 5. Sampel *Authenticity score (Missmatch).*

Review	Sentiment	Rating
Mohon untuk pembayaran QRIS sinkron ke APK DANA DONG MIN.	neutral	positive
hadeuh ga bisa login login	negative	positive
Aplikasi sangat bagus sekali..memudahkan beli bensin..tiap mau transaksi gagal koneksi.barcode spbu tak terbaca..akhirnya malu beli bensin pake apk.aplikasi bumn kok sangat memuaskan banget ya.	positive	negative

Temuan bahwa sekitar 23% data menunjukkan *mismatc*h mengindikasikan bahwa *rating* numerik tidak selalu dapat dijadikan representasi tunggal dari persepsi pengguna. *Rating* bersifat ringkas dan global, sementara analisis berbasis teks khususnya ABSA mampu menangkap nuansa dan fokus pada aspek tertentu dari pengalaman pengguna.

Namun demikian, *mismatch* juga memperlihatkan bahwa model belum sepenuhnya mampu memahami fenomena pragmatik seperti sarkasme atau struktur kalimat dengan kontras makna. Oleh karena itu, ketidaksesuaian tidak hanya mencerminkan keterbatasan *rating*, tetapi juga keterbatasan model dalam memahami kompleksitas bahasa alami.

***Explainable Artificial Intelligence* (XAI)**

Metode ini bekerja dengan menghitung perubahan *output* model ketika representasi suatu token bergerak dari kondisi *baseline* menuju nilai aktualnya, sehingga diperoleh nilai *attribution* yang merepresentasikan tingkat pengaruh token tersebut terhadap kelas tertentu. Dengan pendekatan ini, proses klasifikasi tidak lagi bersifat *black-box*, melainkan dapat ditelusuri hingga ke tingkat kata.

Gambar 4. *Heatmap XAI.*

Pada visualisasi *heatmap* multi-kelas di Gambar 4., setiap token diberikan nilai kontribusi terhadap masing-masing kelas sentimen, yaitu negatif, netral, dan positif. Warna merah menunjukkan bahwa token tersebut memberikan kontribusi positif terhadap kelas tertentu, artinya keberadaan token tersebut meningkatkan probabilitas kelas tersebut dipilih oleh model. Sebaliknya, warna hijau menunjukkan kontribusi negatif, yang berarti token tersebut justru menurunkan probabilitas kelas tersebut. Intensitas warna mencerminkan besarnya pengaruh; semakin pekat warna merah atau hijau, semakin besar dampak token tersebut terhadap keputusan akhir model. Nilai numerik yang lebih tinggi mengindikasikan kontribusi yang lebih dominan, sedangkan nilai yang mendekati nol menunjukkan pengaruh yang relatif kecil. Untuk lebih menjelaskan lagi, dapat dilihat pada Tabel 6.

Tabel 6. Sampel Hasil Transparansi XAI.

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
negative	negative (0.98)	negative	-3.93	[CLS] payment [SEP] metode bayar nya ribet bank sedia [SEP]

Pada sampel yang dianalisis, label sebenarnya adalah negatif dan model memprediksi negatif dengan probabilitas tinggi sebesar 0,98. Token seperti "ribet" memperlihatkan kontribusi yang sangat kuat terhadap kelas negatif, yang ditunjukkan oleh warna merah dengan intensitas tinggi pada baris kelas negatif serta nilai *attribution* yang signifikan. Pada saat yang sama, token tersebut memiliki warna hijau pada kelas netral dan positif, yang berarti keberadaannya secara aktif menurunkan kemungkinan dua kelas tersebut. Pola ini menunjukkan bahwa model secara konsisten mengasosiasikan kata dengan muatan emosional kuat sebagai indikator sentimen negatif. Token lain seperti "metode" dan "bayarnya" juga memberikan kontribusi terhadap kelas negatif, meskipun tidak sebesar kata "ribet", yang menunjukkan bahwa model tidak hanya menangkap ekspresi emosional, tetapi juga memahami konteks aspek pembayaran yang menjadi fokus analisis.

Distribusi kontribusi ini mengindikasikan bahwa keputusan model didasarkan pada fitur linguistik yang relevan dan logis secara semantik. Token yang bersifat netral atau kurang memiliki muatan emosional cenderung memiliki nilai *attribution* kecil dengan warna yang lebih pucat, yang berarti model tidak menjadikannya sebagai faktor utama dalam pengambilan keputusan. Hal ini memperlihatkan bahwa model mampu membedakan antara kata yang bersifat informatif dan kata yang benar-benar menentukan polaritas sentimen.

Di sisi lain, penerapan XAI juga membantu menjelaskan keterbatasan model yang teridentifikasi pada analisis *mismatch* sebelumnya. Pada kalimat yang mengandung sarkasme atau kontras makna, seperti pernyataan yang diawali dengan pujian namun diakhiri dengan kritik, token awal yang bersifat positif dapat memberikan kontribusi besar terhadap kelas positif sebelum konteks lanjutan dipertimbangkan secara menyeluruh. Jika kontribusi token awal tersebut cukup dominan, model dapat terdorong ke arah prediksi yang kurang sesuai dengan maksud keseluruhan kalimat. Dengan demikian, visualisasi *attribution* tidak hanya berfungsi sebagai alat transparansi, tetapi juga sebagai sarana diagnostik untuk memahami bagaimana model memproses struktur kalimat yang kompleks.

Secara keseluruhan, hasil penerapan *Integrated Gradients* menunjukkan bahwa model mengambil keputusan berdasarkan pola linguistik yang dapat dijelaskan dan selaras dengan interpretasi manusia pada kasus-kasus dengan ekspresi sentimen eksplisit. Transparansi ini

meningkatkan tingkat kepercayaan terhadap hasil analisis sentimen berbasis aspek, sekaligus memberikan pemahaman yang lebih mendalam mengenai kekuatan dan keterbatasan model dalam menghadapi variasi bahasa alami. Dengan demikian, teknik *Explainable Artificial Intelligence* dalam penelitian ini berhasil memberikan penjelasan yang terukur dan dapat ditelusuri terhadap proses klasifikasi sentimen, sehingga menjawab rumusan masalah terkait transparansi analisis yang dihasilkan model.

5. KESIMPULAN DAN SARAN

Kesimpulan ditulis secara singkat yaitu mampu menjawab tujuan atau permasalahan penelitian dengan menunjukkan hasil penelitian atau pengujian hipotesis penelitian, tanpa mengulang pembahasan. Kesimpulan ditulis secara kritis, logis, dan jujur berdasarkan fakta hasil penelitian yang ada, serta penuh kehati-hatian apabila terdapat upaya generalisasi. Bagian kesimpulan dan saran ini ditulis dalam bentuk paragraf, tidak menggunakan penomoran atau *bullet.* Pada bagian ini juga dimungkinkan apabila penulis ingin memberikan saran atau rekomendasi tindakan berdasarkan kesimpulan hasil penelitian. Demikian pula, penulis juga sangat disarankan untuk memberikan ulasan terkait keterbatasan penelitian, serta rekomendasi untuk penelitian yang akan datang.

DAFTAR REFERENSI

Aljrees, T., et al. (2024). Contradiction in text review and apps rating: Prediction using textual features and transfer learning. *PeerJ Computer Science, 10*. https://doi.org/10.7717/peerj-cs.1722

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

Ayu, F., Aryanti, D., Luthfiarta, A., Adiwinata, D., & Soeroso, I. (2025). Aspect-based sentiment analysis with LDA and IndoBERT algorithm on mental health app: Riliv. *Journal of Applied Informatics and Computing*. https://doi.org/10.30871/jaic.v9i2.8958

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys, 55*(9). https://doi.org/10.1145/3561048

Fiarni, C., & Cellose, C. J. (2024). Sentiment analysis of Indonesian video streaming application services reviews using fine-tuning IndoBERT and aspect modeling. *Journal of Computer Sciences and Informatics*. https://doi.org/10.5455/JCSI.20240915104407

Jazuli, A., Widowati, & Kusumaningrum, R. (2023). Aspect-based sentiment analysis on student reviews using the Indo-BERT base model. *E3S Web of Conferences, 448*, 02004. https://doi.org/10.1051/e3sconf/202344802004

Liu, B. (2016). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press. https://doi.org/10.1017/CBO9781139084789

Miglani, V., Yang, A., Markosyan, A. H., Garcia-Olano, D., & Kokhlikyan, N. (2023). Using Captum to explain generative language models. https://doi.org/10.18653/v1/2023.nlposs-1.19

Perwira, R. I., Permadi, V. A., Purnamasari, D. I., & Agusdin, R. P. (2025). Domain-specific fine-tuning of IndoBERT for aspect-based sentiment analysis in Indonesian travel user-generated content. *Journal of Information Systems Engineering and Business Intelligence, 11*(1), 30–40. https://doi.org/10.20473/jisebi.11.1.30-40

Selamet, Bejo, A., & Putranto, L. M. (2024). Aspect-based sentiment analysis on PLN Mobile application using word embedding BERT and GRU classification algorithm: A case study of Play Store comments in 2023. In *2024 7th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. IEEE. https://doi.org/10.1109/ISRITI64779.2024.10963459

Tunstall, L., von Werra, L., & Wolf, T. (2022). *Natural language processing with transformers: Building language applications with Hugging Face*. O'Reilly Media.

Wafda, A., Fudholi, D. H., & Nugraha, J. (2025). Aspect-based sentiment analysis on Twitter tweets about the Merdeka Curriculum using IndoBERT. *JITK: Jurnal Ilmu Pengetahuan dan Teknologi Komputer, 10*(3). https://doi.org/10.33480/jitk.v10i3.5692

Yulianti, E., & Nissa, N. K. (2024). ABSA of Indonesian customer reviews using IndoBERT: Single-sentence and sentence-pair classification approaches. *Bulletin of Electrical Engineering and Informatics, 13*(5), 3579–3589. https://doi.org/10.11591/eei.v13i5.8032