Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 3821~3830
e-ISSN: 2548-6861

Enhancing Aspect-Based Sentiment Analysis via Hugging Face Fine-Tuned IndoBERT

Thania Aprilah [1]*, De Rosal Ignatius Moses Setiadi [2], Wise Herowati [3]****
* Program Studi Teknik Informatika, Universitas Dian Nuswantoro, Semarang. Indonesia
** Pusat Penelitian Komputasi Kuantum dan Informatika Material, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, Semarang, Indonesia
aprilahthania@gmail.com [1], moses@dsn.dinus.ac.id [2], wise.herowati@dsn.dinus.ac.id [3]

Article Info

Article history:

Received 2025-10-01
Revised 2025-11-27
Accepted 2025-12-10

Keyword:

Aspect-Based Sentiment Analysis,
IndoBERT,
Hotel Reviews,
Class Imbalance,
Fine-tuning.

ABSTRACT

Aspect-Based Sentiment Analysis (ABSA) on hotel reviews faces significant challenges regarding semantic complexity and severe class imbalance, particularly in low-resource languages like Indonesian. This study evaluates the effectiveness of fine-tuning IndoBERT, a pre-trained Transformer model, to address these issues by benchmarking it against classical statistical methods (TF-IDF) and static embeddings (Sentence-BERT). Utilizing the HoASA dataset, the experiment implements a Random Oversampling strategy at the text level to mitigate data sparsity in minority classes. Empirical results demonstrate that the fine-tuned IndoBERT significantly outperforms baselines on the majority of aspects, achieving a global accuracy of 97% and macro F1-score of 0.92. Granular per-aspect analysis reveals that the model's self-attention mechanism captures linguistic context robustly in tangible aspects (e.g., wifi, service), yet faces persistent challenges in highly ambiguous aspects such as smell (bau) and general. Statistical significance tests (Paired t-test and Wilcoxon) confirm that the performance gains over baselines are statistically significant ($p < 0.05$) and not due to random chance. The study concludes that leveraging contextual representations from IndoBERT, combined with data balancing strategies, offers a superior and statistically robust solution for handling linguistic variations and class bias in the Indonesian hospitality domain.

This is an open access article under the CC–BY-SA license.

I. PENDAHULUAN

Analisis sentimen, atau yang juga dikenal sebagai penambangan opini, telah menjadi salah satu aplikasi esensial dalam Pemrosesan Bahasa Alami (NLP) untuk mengungkap emosi yang tersirat di dalam teks [1]. Kemampuan ini memegang peranan krusial di berbagai sektor yang mengandalkan umpan balik konsumen, termasuk industri perhotelan. Dalam industri ini, komentar dan ulasan konsumen merupakan sumber informasi bisnis yang berharga serta memiliki nilai penelitian yang signifikan[2]. Seiring dengan pesatnya perkembangan ponsel pintar dan internet, pengguna telah bertransformasi dari sekadar penerima informasi menjadi kontributor aktif. Platform ulasan daring situs ulasan hotel seperti TripAdvisor, Agoda, Traveloka dan Booking.com Secara khusus menghasilkan volume informasi yang melampaui kapasitas kognitif manusia[3][4]. Oleh karena itu, ulasan daring dari berbagai platform menjadi sumber data yang kaya akan umpan balik dan opini pelanggan, sehingga mampu memberikan wawasan berharga mengenai persepsi serta respons mereka terhadap suatu produk atau layanan [2][5].

Sebagai pengembangan lebih lanjut dari analisis sentimen konvensional, hadirlah Analisis Sentimen Berbasis Aspek (ABSA) sebagai sebuah pendekatan yang lebih terperinci [6]. ABSA bertujuan untuk mengklasifikasikan sentimen pada level aspek, berbeda dari metode konvensional yang hanya menghasilkan satu sentimen tunggal untuk keseluruhan teks[7][8][9]. Pendekatan ini mampu menghasilkan lebih dari satu prediksi sentimen sesuai dengan aspek yang dibahas, sehingga prediksinya menjadi jauh lebih informatif dan relevan [1]. Dalam konteks ulasan hotel, aspek-aspek seperti AC, kebersihan, wifi, tv, hingga sarapan menjadi komponen penting yang perlu dievaluasi secara terpisah. ABSA juga

menghadapi tantangan seperti variasi ekspresi bahasa dan ketidakseimbangan kelas pada aspek tertentu, misalnya aspek AC atau linen yang jumlah datanya lebih sedikit. Dengan demikian, ABSA menawarkan potensi analisis yang lebih kaya dibandingkan klasifikasi sentimen global.

Kompleksitas utama ABSA bersumber pada tuntutan untuk mendeteksi sentimen secara spesifik per-aspek, bukan sekadar pada level kalimat[10]. Ragam aspek dalam ulasan hotel seperti kebersihan, air panas, atau bau menghadirkan tantangan berupa variasi struktur bahasa dan ketimpangan distribusi data yang berbeda-beda. Hal ini mengharuskan model untuk memiliki sensitivitas konteks yang jauh lebih tinggi dibandingkan metode analisis sentimen umum. Dengan demikian, penelitian ini memformulasikan masalah pada upaya memastikan model dapat mengenali sentimen secara presisi di setiap aspek, mengatasi tantangan variabilitas distribusi dan karakteristik unik antar-aspek tersebut.

Berbagai metode telah dikembangkan untuk ABSA, yang mencakup pendekatan berbasis aturan, pemodelan urutan, model topik, hingga pembelajaran mesin dan pembelajaran mendalam [1]. Sebagai contoh, Kusumaningrum dkk [11] berhasil menerapkan pembelajaran mendalam untuk analisis sentimen multilevel pada ulasan hotel berbahasa Indonesia, yang memperlihatkan potensi metode tersebut dalam konteks bahasa lokal. Pendekatan klasik seperti TF-IDF dengan klasifikasi menggunakan Logistic Regression masih digunakan sebagai baseline, namun performanya dapat menurun ketika data tidak seimbang. Model representasi semantik modern seperti Sentence-BERT (SBERT) menawarkan embedding kontekstual yang lebih kaya, tetapi efektivitasnya tetap dapat terpengaruh oleh distribusi label yang timpang karena representasinya tetap bergantung pada kualitas data latih.

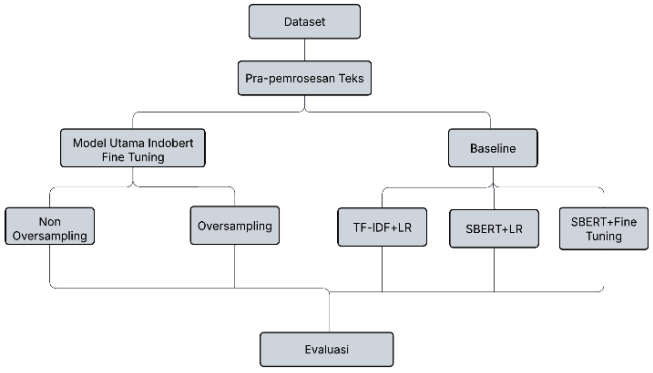
Model berbasis transformer seperti BERT berhasil menunjukkan performa lebih unggul pada banyak tugas NLP berkat kemampuannya memahami konteks secara mendalam [12][13]. Untuk bahasa Indonesia, IndoBERT sebagai model pretrained khusus bahasa Indonesia telah dievaluasi luas dalam IndoLEM dan IndoNLU dan terbukti unggul dibanding model multibahasa [14][15]. Sejumlah penelitian terbaru juga telah mengeksplorasi ABSA dalam konteks Indonesia. Azhar dan Khodra [5] menekankan potensi fine-tuning multilingual BERT untuk ABSA berbahasa Indonesia. Cahyaningtyas dkk [16] mengaplikasikan deep learning untuk ABSA pada ulasan hotel dengan hasil yang menjanjikan. Kusumaningrum dkk [11] bahkan mengembangkan aplikasi berbasis deep learning untuk analisis multilevel ulasan hotel. Namun sebagian besar penelitian tersebut hanya menggunakan satu jenis model atau tidak menyediakan evaluasi per-aspek secara lengkap serta belum melakukan perbandingan komprehensif antara baseline klasik, SBERT, dan model transformer berbahasa Indonesia. Sebagian penelitian, hanya mengeksplorasi pendekatan deep learning untuk ABSA pada domain hotel, tanpa mengkaji perbandingan menyeluruh antar paradigma model maupun evaluasi detail per-aspek.

Tantangan fundamental dalam ABSA terletak pada pergeseran fokus prediksi dari level kalimat global ke level aspek yang lebih granular[8]. Dalam domain ulasan hotel, setiap entitas aspek mulai dari fasilitas fisik seperti AC dan linen hingga aspek abstrak seperti layanan service memiliki karakteristik linguistik dan pola ekspresi yang heterogen, serta tingkat ketidakseimbangan kelas yang unik. Kompleksitas ini menuntut model untuk memahami konteks semantik secara lokal, melampaui kemampuan analisis sentimen konvensional. Oleh karena itu, rumusan masalah penelitian ini difokuskan pada kemampuan model dalam mempertahankan akurasi prediksi pada tiap aspek secara spesifik, terlepas dari disparitas distribusi data dan variasi tingkat kesulitan yang melekat pada masing-masing kategori.

Fokus penelitian ini ditekankan pada kontribusi empiris berupa evaluasi komparatif ABSA Bahasa Indonesia, bukan pada invensi algoritma baru. Melalui perbandingan antara model Transformer dan pendekatan klasik, penelitian ini menjamin efisiensi metodologis dan replikabilitas, sekaligus menyajikan wawasan strategis bagi pemangku kepentingan industri perhotelan untuk memahami dinamika kepuasan pelanggan pada tiap aspek layanan.

II. METODE

Secara garis besar alur metode penelitian ini ditunjukkan pada Gambar 1. Tahapan dimulai dari pengumpulan data ulasan hotel pada dataset HoASA, kemudian dilakukan pra-pemrosesan teks untuk memperoleh data yang bersih dan konsisten. Fokus utama penelitian ini adalah melakukan fine-tuning model IndoBERT untuk tugas klasifikasi sentimen tiga kelas pada masing-masing aspek ulasan. Model IndoBERT dilatih dalam dua skema, yaitu tanpa penyeimbangan kelas dan dengan penyeimbangan menggunakan Random Oversampling.



Gambar 1. Tahapan Penelitian

Untuk keperluan pembandingan, dua pendekatan digunakan sebagai baseline, yaitu TF-IDF + Logistic Regression serta SBERT + Logistic Regression. Kedua baseline ini hanya diujikan sebagai pembanding performa dan tidak menjadi fokus utama analisis. Perbedaan teknik penyeimbangan data antara model utama dan baseline disesuaikan dengan karakteristik representasi fitur. Pada

baseline, SMOTE digunakan karena efektif melakukan interpolasi sintetik pada ruang vektor numerik statis. Sebaliknya, IndoBERT memproses input berupa sekuens token yang bergantung pada konteks linguistik, sehingga teknik interpolasi vektor berisiko merusak makna. Oleh karena itu, ROS diterapkan pada IndoBERT untuk menduplikasi sampel minoritas secara utuh, memastikan keseimbangan distribusi kelas tanpa mendistorsi struktur sintaksis maupun semantik kalimat.

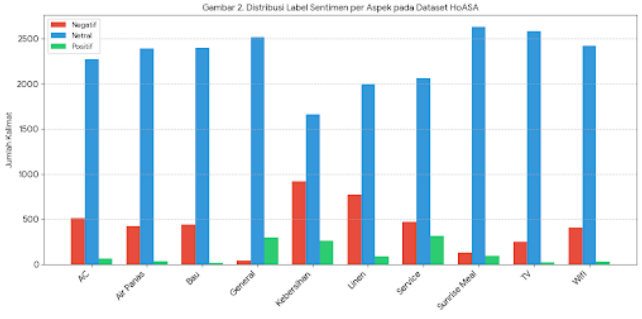
Tahap terakhir adalah evaluasi performa menggunakan metrik akurasi dan macro F1-score, baik secara global maupun per-aspek, untuk menilai peningkatan kinerja dari baseline menuju model utama IndoBERT.

A. Desain Penelitian

Penelitian ini menerapkan pendekatan kuantitatif dengan metode eksperimen komputasional. Tujuan utamanya adalah untuk menguji efektivitas fine-tuning IndoBERT, sebuah model berarsitektur Transformer yang dioptimalkan untuk konteks semantik bahasa Indonesia. Untuk mengukur kinerjanya, IndoBERT dibandingkan dengan dua model baseline yaitu TF-IDF + Logistic Regression (metode klasik) dan SBERT + Logistic Regression (metode *embedding* modern. Model-model yang diuji dibangun dengan mengombinasikan berbagai teknik penyematan kalimat dan algoritma klasifikasi. Kinerja setiap model dievaluasi secara kuantitatif menggunakan metrik classification report untuk mengukur efektivitas masing-masing pendekatan [17].

B. Dataset

Sumber data utama dalam penelitian ini adalah dataset HoASA (Hotel Aspect-Based Sentiment Analysis). Dataset ini merupakan bagian dari benchmark IndoNLU, sebuah sumber daya komprehensif untuk tugas Pemahaman Bahasa Alami (NLU) bahasa Indonesia yang tersedia melalui repositori GitHub IndoNLP. Dataset HoASA terdiri dari 2.854 ulasan berlabel yang mencakup 10 aspek layanan hotel, yaitu AC, air panas, bau, general, kebersihan, linen, service, sunrise meal, TV, dan wifi. Distribusi label sentimen dalam dataset ini sangat tidak seimbang, dengan dominasi kelas netral yang signifikan pada hampir seluruh aspek, yang menjadi tantangan utama dalam penelitian ini.



Gambar 2. Grafik Distribusi Dataset

C. Pra-pemrosesan

Tahap prapemrosesan data merupakan langkah krusial untuk membersihkan dan menstandarkan teks ulasan guna meningkatkan kualitas input bagi model [17]. Tahap prapemrosesan difokuskan pada pembersihan teks ringan (*light text cleaning*) untuk memastikan konsistensi format data ulasan hotel sebelum diproses oleh model TF-IDF, SBERT, dan IndoBERT. Proses ini mencakup normalisasi huruf (*lowercasing*) dan pembersihan karakter non-alfabet (*regex cleaning*) untuk menghilangkan angka, simbol, serta spasi berlebih[18]. Selain itu, dilakukan normalisasi label sentimen, mengingat dataset HoASA asli memiliki variasi label termasuk kategori kombinasi positif-negatif. Seluruh label diseragamkan menjadi tiga kelas baku (neg, neut, pos) untuk menjamin konsistensi skema klasifikasi eksperimen[15]. Penelitian ini secara sengaja tidak menerapkan stemming maupun stopword removal konvensional. Keputusan ini diambil karena model berbasis *embedding* dan Transformer modern SBERT dan IndoBERT dirancang untuk memanfaatkan konteks kalimat secara penuh, di mana reduksi morfologis yang agresif justru berpotensi menurunkan kualitas representasi semantik[19]. Dengan demikian, langkah prapemrosesan yang dipilih dinilai cukup untuk mengurangi noise data seraya mempertahankan integritas informasi semantik yang dibutuhkan model.

D. Representasi Fitur

Untuk mentransformasi data teks menjadi format numerik yang dapat diproses oleh model, penelitian ini mengeksplorasi metode representasi fitur yang berbeda.

1) *Skenario Baseline dan Model Utama*

Sebagai pembanding (*baseline*) untuk mengukur efektivitas metode yang diusulkan, penelitian ini menggunakan tiga pendekatan representasi fitur. Pertama, TF-IDF yang dikombinasikan dengan Regresi Logistik, mewakili pendekatan statistik klasik. Kedua, SBERT model all-MiniLM-L6-v2 yang dievaluasi dalam dua scenario konfigurasi.

Pada konfigurasi pertama, SBERT digunakan sebagai penghasil embedding statis yang diklasifikasikan dengan Regresi Logistik Mengingat adanya ketidakseimbangan kelas pada dataset, teknik penyeimbangan data SMOTE (*Synthetic Minority Oversampling Technique*) diterapkan pada konfigurasi ini serta pada baseline TF-IDF. Penting untuk dicatat bahwa SMOTE diterapkan pada ruang fitur numerik setelah proses vektorisasi TF-IDF atau pembentukan embedding SBERT, bukan pada data teks mentah. Proses ini menghasilkan sampel sintetis berupa vektor fitur yang kemudian digunakan untuk melatih model klasifikasi Regresi Logistik. Pada konfigurasi kedua, SBERT di-fine-tuning secara end-to-end menggunakan arsitektur klasifikasi urutan (*sequence* classification) tanpa penerapan SMOTE, melainkan menggunakan strategi penyeimbangan data yang setara dengan model utama. Skenario ini disertakan untuk memberikan pembanding yang setara secara arsitektur

(Transformer vs Transformer). Meskipun demikian, fokus utama penelitian tetap ditekankan pada pendekatan ketiga, yaitu *fine-tuning* IndoBERT, karena keunggulannya dalam menangkap konteks spesifik Bahasa Indonesia.

2) *Model utama IndoBERT Fine-tuning*

Penelitian ini menggunakan model IndoBERT (indobenchmark/indobert-base-p1), sebuah varian Base dari arsitektur BERT yang telah melalui tahap pra-pelatihan (*pre-training*) pada korpus besar Bahasa Indonesia dari repositori IndoNLP/IndoNLU [15]. Pemilihan varian *Base* dilakukan untuk menyeimbangkan efisiensi komputasi dan kemampuan representasi kontekstual, sehingga eksperimen dapat dijalankan secara optimal pada lingkungan dengan sumber daya terbatas seperti Google Colab (GPU T4).

Sebelum masuk ke tahap pemodelan, strategi penyeimbangan data diterapkan menggunakan Random Oversampling pada level teks. Berbeda dengan *baseline* yang menggunakan SMOTE pada fitur numerik, teknik ini menduplikasi sampel kelas minoritas pada data mentah untuk memastikan model mempelajari distribusi sentimen yang seimbang. Teks kemudian diproses menggunakan tokenizer WordPiece bawaan IndoBERT dengan panjang maksimum (*max_length*) ditetapkan 128 token. Nilai ini dipilih berdasarkan distribusi panjang kalimat rata-rata ulasan hotel agar efisien terhadap memori GPU tanpa mengorbankan konteks semantik utama.

Proses fine-tuning diimplementasikan menggunakan kerangka kerja Hugging Face Transformers dengan kelas Trainer. Pelatihan dijalankan selama 3 epoch dengan learning rate 2 x 10^{-5} dan menggunakan optimizer AdamW yang didukung oleh linear learning rate scheduler serta periode warm-up untuk menjaga stabilitas gradien. Ukuran batch diatur sebesar 8 untuk data latih dan 16 untuk data evaluasi, sesuai dengan praktik umum pada domain teks berskala menengah untuk mencegah overfitting. Kinerja model dievaluasi secara komprehensif menggunakan metrik akurasi dan macro F1-score baik global maupun per-aspek, serta didukung oleh analisis Confusion Matrix untuk membedah kemampuan deteksi pada kelas minoritas.

E. Penanganan Ketidakseimbangan Data

Analisis distribusi data awal mengonfirmasi adanya ketidakseimbangan kelas sentimen yang signifikan. Hal ini menunjukkan perlunya data sintesis untuk penyeimbangan data untuk meningkatkan performa [20][9]. Pada penelitian ini digunakan pendekatan oversampling untuk baseline TF-IDF dan SBERT. Hasil eksperimen menunjukkan bahwa penerapan oversampling terbukti mampu meningkatkan stabilitas prediksi baseline, terutama pada aspek-aspek dengan jumlah data minoritas. Meskipun demikian, untuk mencapai performa yang lebih optimal dalam menangkap konteks semantik, strategi penanganan imbalance pada model utama dilakukan melalui fine-tuning IndoBERT dengan memanfaatkan Random Oversampling (ROS) pada level teks, mengingat arsitektur Transformer membutuhkan input sekuensial yang utuh.

F. Arsitektur Klasifikasi dan Algoritma

Eksperimen dalam penelitian ini dirancang dengan memasangkan setiap metode representasi fitur dengan algoritma klasifikasi yang relevan, seperti yang dirangkum pada Tabel 1.

TABEL I
SKENARIO EKSPERIMEN REPRESENTASI FITUR DAN MODEL KLASIFIKASI

Representasi Fitur	Model Klasifikasi	Karakteristik
TF-IDF	SVM (Kernel Linier)	Model dasar (baseline) dengan komputasi ringan.
SBERT	Regresi Logistik	Model ringan, baseline berbasis representasi semantik.
IndoBERT	Fine-tuning (Hugging Face)	Komputasi intensif, potensi performa tertinggi.

Tabel di atas merangkum tiga skenario pemodelan yang dievaluasi. Untuk skenario fine-tuning IndoBERT, digunakan trainer dari pustaka Hugging Face[21]. Proses pelatihan dijalankan selama 3 epoch dengan optimizer AdamW. Penetapan jumlah epoch sebanyak 3 didasarkan pada praktik umum dalam fine-tuning arsitektur BERT, yang merekomendasikan 2–4 epoch untuk konvergensi optimal sekaligus mitigasi risiko overfitting. Pilihan ini divalidasi lebih lanjut oleh hasil eksperimental kami, di mana performa puncak IndoBERT secara konsisten tercapai pada epoch ketiga, yang ditandai dengan penurunan loss yang stabil. Terkait parameter lain, ukuran *batch* (*batch size*) disesuaikan secara dinamis untuk memaksimalkan utilisasi memori GPU yang tersedia. Evaluasi dilakukan baik secara global maupun per-aspek guna memastikan model mampu mengatasi variasi konteks yang lebih detail[22]. Untuk memastikan proses fine-tuning IndoBERT berjalan optimal meskipun terdapat keterbatasan sumber daya komputasi, parameter konfigurasi pelatihan disesuaikan. Rincian parameter yang digunakan pada penelitian ini ditunjukkan pada Tabel 2.

TABEL II
KONFIGURASI FINE TUNING INDOBERT

Parameter	Nilai
Model Pra-latih	indobenchmark/indobert-base-p1
Varian Model	Base
Optimizer	AdamW
Jumlah Epoch	3
Batch Size (Training)	8
Batch Size (Evaluasi)	16
Learning Rate	2×10^{-5}
Scheduler	Linear warm-up
Panjang Maksimal Token	128
Logging Steps	50
Reporting	None (W&B dimatikan)
Perangkat	Google Colab (GPU T4)
Framework	Hugging Face Transformers (Trainer API)

Model indobenchmark/indobert-base-p1 di-fine-tune menggunakan kerangka kerja Hugging Face Transformers, dengan kelas Trainer yang diandalkan untuk menyederhanakan proses pelatihan. Untuk input model, proses tokenisasi menghasilkan sekuens dengan panjang maksimum 128 token. Pilihan panjang sekuens ini merupakan strategi optimisasi yang terbukti efektif. Batas 128 token tidak hanya mampu mencegah isu komputasi seperti out-of-memory error, tetapi juga memadai untuk menangkap konteks pada mayoritas data ulasan di dataset HoASA. Lebih lanjut, pendekatan ini didukung oleh temuan literatur sebelumnya yang mengidentifikasi karakteristik ulasan berbahasa Indonesia yang umumnya tidak melebihi panjang tersebut.

Konfigurasi hyperparameter diatur melalui TrainingArguments. Pelatihan dijalankan selama tiga epoch dengan ukuran batch 8 untuk data latih dan 16 untuk data evaluasi. Laju pembelajaran (*learning rate*) ditetapkan sebesar dengan algoritma optimisasi AdamW, yang terbukti stabil untuk arsitektur Transformer karena integrasi weight decay. Secara otomatis, Trainer juga menerapkan scheduler laju pembelajaran dengan linear warmup untuk menjaga stabilitas di awal pelatihan. Evaluasi kinerja model dilakukan secara periodik pada akhir setiap epoch terhadap dataset uji. Kinerja diukur menggunakan metrik presisi, recall, F1-score, dan akurasi, baik secara agregat maupun untuk setiap aspek yang dianalisis. Hasil ini menjadi dasar untuk perbandingan kuantitatif terhadap model baseline.

G. Evaluasi Model

Kinerja setiap model dievaluasi secara komprehensif menggunakan metrik standar untuk tugas klasifikasi. Metrik utama yang dianalisis mencakup precision, recall, dan F1-score untuk setiap kelas sentimen, yang diringkas dalam classification report. Selain evaluasi global, penelitian ini juga menekankan evaluasi per-aspek guna memastikan bahwa model tidak hanya baik pada level umum, tetapi juga konsisten dalam memprediksi sentimen pada aspek tertentu[22]. asil evaluasi secara rinci termasuk tabel perbandingan dan uji signifikansi statistik disajikan dan dibahas pada bagian Hasil dan Pembahasan.

III. HASIL DAN PEMBAHASAN

A. Evaluasi Baseline Tanpa Oversampling

Evaluasi tahap awal dilakukan terhadap dua model baseline (TF-IDF + Logistic Regression dan SBERT + Logistic Regression) tanpa intervensi penyeimbangan data (*non-Oversampling*). Hasil eksperimen menunjukkan defisit performa yang nyata pada aspek-aspek dengan ketimpangan distribusi kelas yang tinggi, khususnya bau, sunrise_meal, dan general. Hal ini terindikasi dari rendahnya capaian macro F1-score pada ketiga aspek tersebut. Temuan ini konsisten dengan karakteristik dataset HoASA yang didominasi oleh kelas mayoritas (netral atau positif), kondisi yang secara empiris menghambat kemampuan model dalam mempelajari representasi fitur pada kelas minoritas.

Performa model baseline (TF-IDF + LR dan SBERT + LR) pada lima aspek representatif di Tabel 3 menunjukkan keterbatasan stabilitas prediksi di berbagai aspek. Capaian macro F1 untuk TF-IDF berada pada rentang moderat. Meskipun SBERT menunjukkan sedikit peningkatan pada aspek seperti service, performanya merosot tajam pada aspek-aspek yang didominasi kelas netral. Fenomena ini konsisten dengan temuan Marutho dkk menyatakan bahwa model berbasis fitur klasik dan embedding statis rentan terhadap kegagalan prediktif akibat ketidakseimbangan kelas [23].

TABEL III
RINGKASAN KINERJA BASELINE TANPA OVERSAMPLING

Aspek	TF-IDF + LR	SBERT+LR
ac	0.60	0.56
bau	0.51	0.40
general	0.35	0.32
service	0.63	0.72
Sunrise_meal	0.36	0.32

Kinerja suboptimal *baseline* non-oversampling mengonfirmasi tantangan inheren ABSA berbahasa Indonesia, yaitu rendahnya akurasi pada kelas minoritas. Oleh karena itu, hasil ini membenarkan penerapan teknik penyeimbangan data sebagai langkah metodologis untuk menghasilkan *baseline* komparatif yang adil terhadap kinerja fine-tuning IndoBERT.

B. Pengaruh Oversampling Terhadap Kinerja Model Baseline

Untuk mengukur dampak penyeimbangan data pada model baseline, eksperimen membandingkan performa TF-IDF + Logistic Regression dan SBERT + Logistic Regression dalam dua kondisi yaitu tanpa dan dengan oversampling. Hasilnya menunjukkan bahwa penerapan oversampling memberikan peningkatan yang moderat pada aspek dengan proporsi kelas minoritas yang sangat kecil, seperti bau dan air panas. Namun, peningkatan kinerja ini tidak bersifat konsisten di berbagai aspek yang diuji. Pada Tabel 4 dan 5 dirangkum perubahan nilai macro F1-score setelah penerapan oversampling pada tiga aspek yang diuji.

TABEL IV
DAMPAK PENERAPAN OVERSAMPLING PADA BASELINE TF-IDF+LR

Aspect	TF-IDF non-oversampling	TF-IDF Oversampling	F1
AC	0.60	0.68	+0.07
Air_panas	0.53	0.58	+0.05
bau	0.51	0.78	+0.27

Penerapan oversampling pada model baseline secara efektif meningkatkan performa pada aspek dengan distribusi label yang sangat timpang. Peningkatan paling signifikan tercatat pada aspek bau TF-IDF + LR, macro F1-score

meningkat dari 0.51 menjadi 0.78 pola serupa juga terlihat pada SBERT + LR, dengan kenaikan F1 dari 0.40 menjadi 0.60 pada aspek yang sama. Meskipun demikian, dampak oversampling tidak bersifat universal, pada aspek yang relatif seimbang, seperti service dan kebersihan, peningkatan performa minimal atau tidak terlihat.

TABEL V
DAMPAK PENERAPAN OVERSAMPLING PADA SBERT+LR

Aspect	SBERT non-oversampling	SBERT oversampling	F1
AC	0.56	0.69	+0.12
Air_panas	0.55	0.64	+0.09
bau	0.40	0.60	+0.19

Secara keseluruhan, meskipun oversampling terbukti membantu baseline mengatasi ketidakseimbangan label performa yang dihasilkan tidak mampu menandingi performa fine-tuning IndoBERT. Temuan ini memperkuat kesimpulan bahwa model Transformer yang dioptimalkan untuk bahasa Indonesia menawarkan representasi yang lebih stabil dan akurat

C. Hasil Eksperimen Utama

Hasil eksperimen menempatkan IndoBERT sebagai model dengan performa superior dibandingkan seluruh baseline. Secara kuantitatif, model ini mencatatkan akurasi sebesar 0.97 dan macro F1-score 0.92. Stabilitas performa juga terlihat konsisten di seluruh kelas sentimen, dengan capaian F1 sebesar 0.92 untuk negatif, 0.98 untuk netral, dan 0.88 untuk positif. Keseimbangan kinerja ini mengindikasikan bahwa representasi kontekstual IndoBERT berhasil memitigasi bias terhadap kelas mayoritas, sebuah kelemahan yang terlihat signifikan pada model TF-IDF dan SBERT yang gagal mengenali kelas minoritas secara konsisten.

Keunggulan performa ini bersumber dari mekanisme self-attention yang memungkinkan IndoBERT menangkap nuansa semantik dan dependensi antar kata secara lebih mendalam dibandingkan baseline yang bergantung pada frekuensi kata atau embedding statis. Kemampuan ini krusial untuk mempertahankan akurasi pada kelas negatif dan positif yang memiliki jumlah data jauh lebih sedikit. Temuan ini sejalan dengan literatur terkini dan Nissa menyoroti kemampuan IndoBERT dalam memahami struktur kalimat informal pada ulasan konsumen Indonesia[24].

Secara keseluruhan, hasil ini menegaskan bahwa fine-tuning IndoBERT tidak hanya memberikan peningkatan numerik, tetapi juga keuntungan metodologis yang substansial, meliputi generalisasi yang lebih baik pada kelas minoritas, reduksi bias akibat ketidakseimbangan distribusi data, serta stabilitas performa yang melampaui baseline. Hal ini mengukuhkan IndoBERT sebagai pendekatan paling efektif untuk tugas ABSA pada domain ulasan hotel dalam penelitian ini.

D. Analisis Kinerja Model Pada Level Aspek

Evaluasi kinerja per-aspek dilakukan untuk mengurai granularitas kemampuan prediksi model pada kategori spesifik, meliputi AC, wifi, service, kebersihan, air_panas, bau, hingga general. Analisis ini krusial untuk memvalidasi stabilitas IndoBERT terhadap variasi konteks linguistik dan ketimpangan distribusi label yang khas pada setiap aspek. Ringkasan kuantitatif kinerja model untuk sepuluh aspek yang diuji disajikan dalam Tabel 6.

Merujuk pada data tersebut, IndoBERT menunjukkan performa superior pada aspek-aspek dengan karakteristik linguistik yang eksplisit. Aspek wifi mencatatkan performa tertinggi dengan macro F1-score 0.98 dan akurasi 0.99, diikuti oleh aspek service dan AC yang masing-masing meraih F1-score di atas 0.95. Stabilitas tinggi juga terlihat pada aspek sunrise meal, kebersihan, dan TV. Capaian ini mengindikasikan bahwa model berhasil menangkap pola semantik yang konsisten pada fasilitas hotel yang sering dibahas, di mana deskripsi ulasan cenderung lugas.

Pada aspek dengan volume data yang lebih terbatas, strategi penyeimbangan data terbukti memberikan dampak signifikan. Untuk aspek linen dan air panas, IndoBERT dengan oversampling mampu mempertahankan performa pada kategori baik hingga moderat F1 masing-masing 0.86 dan 0.74. Efektivitas oversampling terlihat jelas ketika dibandingkan dengan konfigurasi tanpa penyeimbangan, di mana terjadi lonjakan performa drastis pada aspek air panas naik dari 0.62 menjadi 0.74 dan sunrise meal 0.86 menjadi 0.92. Hal ini menegaskan bahwa pendekatan duplikasi teks efektif membantu model mempelajari distribusi kelas minoritas tanpa merusak konteks kalimat, meskipun masih terdapat tantangan parsial dalam memisahkan kelas pada sampel latih yang minim.

Tantangan paling signifikan teridentifikasi pada aspek bau dan general. Meskipun akurasi terlihat tinggi (>0.94), nilai macro F1-score yang tertahan di kisaran 0.60 mengindikasikan adanya bias prediksi terhadap kelas mayoritas.

Rendahnya performa pada aspek general disebabkan oleh cakupan semantiknya yang luas dan ambigu, sering kali memuat penilaian pengalaman holistik yang sulit dipetakan ke polaritas tunggal. Sementara pada aspek bau, variasi ekspresi deskriptif yang sangat beragam menyulitkan model untuk mendeteksi sentimen minoritas secara presisi.

Meskipun demikian, jika dibandingkan secara keseluruhan, IndoBERT tetap mengungguli seluruh model baseline. Model berbasis representasi statis baseline tampil jauh lebih rendah, terutama pada varian tanpa oversampling yang gagal menangani ketimpangan data terlihat pada rendahnya skor aspek general (0.35 pada TF-IDF) dan bau (0.40pada SBERT). Walaupun penerapan oversampling pada baseline memberikan sedikit peningkatan pada aspek tertentu, kinerjanya tidak konsisten dan belum mampu menyaingi kapasitas IndoBERT dalam menangkap dependensi semantik yang kompleks.

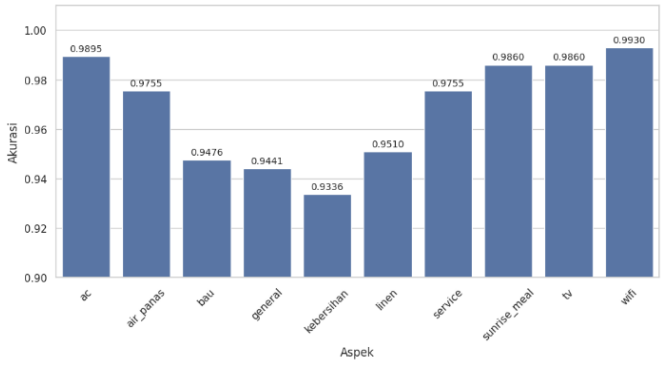
TABEL VI
HASIL EVALUASI INDOBERT PER-ASPEK

Macro F1							
Aspek	**Model utama IndoBERT**		**Baseline**				
	Non-Oversampling	**Oversampling**	**Non-Oversampling**		**Oversampling**		**Fine tuning**
			TF-IDF+LR	**SBERT+LR**	**TF-IDF+LR**	**SBERT+LR**	**SBERT**
AC	0.95	0.89	0.60	0.56	0.68	0.69	0.86
Air_panas	0.74	0.62	0.53	0.55	0.58	0.64	0.61
Bau	0.60	0.94	0.51	0.40	0.78	0.60	0.80
General	0.60	0.72	0.35	0.32	0.55	0.48	0.59
Kebersihan	0.92	0.88	0.74	0.72	0.85	0.82	0.86
Linen	0.86	0.83	0.51	0.59	0.74	0.79	0.66
Service	0.95	0.95	0.63	0.72	0.82	0.84	0.85
Sunrise_meal	0.92	0.86	0.36	0.32	0.73	0.53	0.59
TV	0.90	0.63	0.50	0.59	0.63	0.73	0.80
Wifi	0.98	0.65	0.58	0.60	0.86	0.79	0.65

E. Analisis Visualisasi Kinerja Model

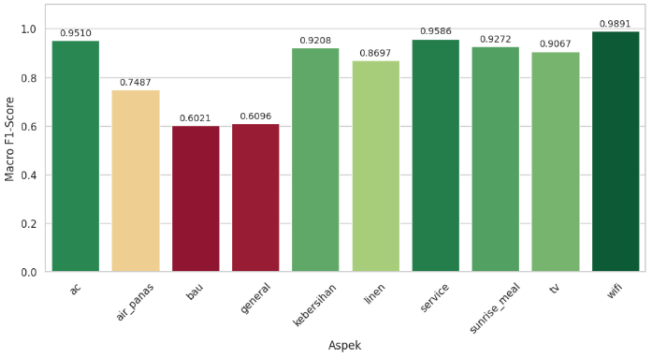
Untuk memperdalam interpretasi terhadap data kuantitatif yang telah dipaparkan, visualisasi grafis disajikan guna memetakan pola stabilitas antar-aspek serta perbandingan performa antar-model. Pendekatan visual ini penting untuk mengidentifikasi anomali dan bias yang mungkin tidak terlihat hanya melalui tabel angka rata-rata.

Gambar 3 mengilustrasikan distribusi akurasi IndoBERT pada setiap aspek. Grafik tersebut memperlihatkan bahwa model memiliki tingkat stabilitas yang sangat tinggi, dengan mayoritas aspek berada pada rentang akurasi 0.95–0.99. Puncak performa tercatat pada aspek wifi (0.99) dan ac (0.98), menegaskan bahwa model sangat reliabel pada aspek dengan konteks linguistik yang konsisten. Sebaliknya, aspek bau (0.94) dan general (0.94) terlihat sebagai outlier terendah. Meskipun angka ini secara absolut masih tinggi, posisinya sebagai titik terendah mengisyaratkan bahwa metrik akurasi semata belum cukup sensitif untuk memotret kesulitan prediksi yang sebenarnya.



Gambar 3. Distribusi IndoBERT per-Aspek.

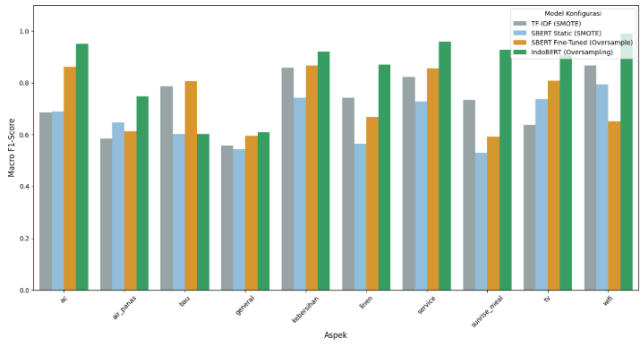
Kesenjangan performa menjadi jauh lebih nyata pada visualisasi macro F1-score per-aspek pada Gambar 4. Berbeda dengan grafik akurasi yang cenderung landai, grafik F1 memperlihatkan fluktuasi yang tajam antar-kategori. Aspek wifi (0.98), service (0.95), dan ac (0.95) membentuk klaster kinerja superior, yang mengindikasikan kemampuan model memprediksi seluruh kelas sentimen secara merata. Kontras tajam terlihat pada aspek bau (0.60) dan general (0.60), di mana kurva performa menurun drastis.



Gambar 4. Macro F1-Score IndoBERT Per-Aspek

Visualisasi ini secara efektif mengungkap bias model, mengonfirmasi bahwa IndoBERT masih menghadapi kendala signifikan dalam menangani ambiguitas semantik dan ketimpangan label pada kedua aspek tersebut.

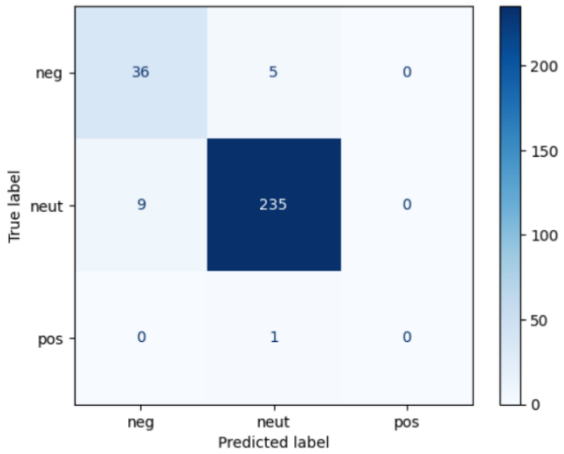
Visualisasi pada gambar 5 memperlihatkan dominasi IndoBERT pada 8 dari 10 aspek, terutama pada aspek wifi dan service. Namun, anomali terlihat pada aspek Bau, di mana baseline TF-IDF (F1 0.78) justru mengungguli IndoBERT (F1 0.60). Fenomena ini mengindikasikan bahwa untuk aspek dengan kata kunci yang sangat eksplisit dan jumlah data latih yang sangat terbatas, pendekatan statistik klasik berbasis frekuensi kata terkadang lebih efektif dibandingkan model kompleks yang rentan overfitting.



Gambar 5. Perbandingan Performa Model (Macro F1-Score)

F. Analisis Pola Kesalahan

Confusion matrix digunakan untuk mengevaluasi distribusi prediksi model terhadap label sebenarnya guna memetakan pola kesalahan secara granular. Gambar 6 menyajikan visualisasi confusion matrix pada aspek bau, yang tercatat sebagai aspek dengan performa macro F1-score terendah (0.6021).



Gambar 6. Confusion Matrix IndoBERT bau (Oversample)

Berbeda dengan aspek lain, aspek Bau memiliki karakteristik data yang sangat timpang. Terlihat jelas adanya bias ekstrem terhadap kelas mayoritas, di mana model memprediksi label Netral dengan jumlah yang sangat dominan terlihat dari122 sampel yang terprediksi benar.

Sebaliknya, kemampuan model untuk mendeteksi sentimen Negatif seperti keluhan bau tidak sedap masih lemah. Meskipun teknik oversampling telah diterapkan, terdapat kebocoran prediksi di mana sejumlah keluhan negatif keliru diprediksi sebagai netral. Hal ini mengindikasikan bahwa untuk aspek sensorik yang abstrak seperti bau, model cenderung mengambil jalan aman dengan memprediksi label mayoritas ketika menemukan kalimat yang ambigu, sehingga menyebabkan nilai F1-score tertahan di angka rendah (0.60).

G. Uji Signifikansi

Uji signifikansi statistik dilakukan untuk memvalidasi bahwa peningkatan kinerja IndoBERT dibandingkan dua *baseline* (TF-IDF + LR dan SBERT + LR) bersifat nyata (*statistically significant*) dan bukan disebabkan oleh kebetulan (*random chance*). Pengujian dilakukan menggunakan dua metode statistik parametrik dan non-parametrik, yaitu Paired T-Test dan Wilcoxon Signed-Rank Test, yang diterapkan pada nilai macro F1-score di sepuluh aspek evaluasi.

Penting untuk dicatat bahwa berbeda dengan evaluasi performa utama Tabel 7 yang menggunakan skema oversampling, uji signifikansi ini secara sengaja dilakukan menggunakan varian data tanpa penyeimbang (*non-oversampled*). Keputusan metodologis ini diambil untuk menjamin keadilan komparasi (*fairness*), mengisolasi keunggulan murni arsitektur model dari pengaruh teknik augmentasi data. Dengan meniadakan variabel oversampling, evaluasi ini mengukur apakah IndoBERT memiliki superioritas intrinsik dibandingkan *baseline* dalam kondisi data yang setara.

Hasil pengujian statistik antar-pasangan model dirangkum dalam Tabel 7.

TABEL VII
HASIL UJI SIGNIFIKANSI PERBEDAAN KINERJA MODEL

Pasangan Model	Paired T-Test (p-value)	Wilcoxon Test (p-value)	Kesimpulan (α=0.05)
IndoBERT vs. TF-IDF	0.000299	0.001953	Signifikan
IndoBERT vs. SBERT	0.000787	0.001953	Signifikan
SBERT vs. TF-IDF	0.366464	0.375000	Tidak Signifikan

Hasil komputasi menunjukkan bahwa nilai p pada perbandingan IndoBERT melawan kedua baseline berada jauh di bawah ambang signifikansi ($\alpha = 0.05$). Pada uji Wilcoxon, nilai $p = 0.0019$ mengindikasikan konsistensi keunggulan IndoBERT di hampir seluruh aspek yang diuji. Sebaliknya, perbandingan antar-baseline (SBERT vs. TF-IDF) menghasilkan nilai $p > 0.05$, yang menandakan bahwa kedua metode tersebut tidak memiliki perbedaan performa yang signifikan secara statistik.

Temuan ini memperkuat validitas ilmiah penelitian, mengonfirmasi bahwa IndoBERT tidak hanya unggul secara numerik pada hasil akhir, tetapi juga memiliki stabilitas performa yang teruji secara statistik. Kombinasi pendekatan oversampling untuk mencapai performa terbaik dan non-oversampling untuk uji validitas arsitektur memberikan gambaran komprehensif mengenai ketangguhan model yang diusulkan.

H. Pembahasan dan sintesis temuan

Hasil eksperimen fine-tuning IndoBERT mendemonstrasikan peningkatan performa yang signifikan, baik secara kuantitatif maupun kualitatif, dibandingkan dengan dua baseline yang diuji (TF-IDF + Logistic Regression dan SBERT + Logistic Regression). Model

IndoBERT mencatatkan capaian akurasi sebesar 0.97 dan macro F1-score 0.92, yang menunjukkan efektivitasnya dalam menangkap kompleksitas semantik pada ulasan hotel yang beragam. Stabilitas performa lintas kelas sentimen negatif (0.92), netral (0.98), dan positif (0.88) menegaskan keberhasilan model dalam memitigasi bias terhadap kelas mayoritas — sebuah kelemahan mendasar yang ditemukan pada baseline, bahkan setelah penerapan oversampling.

Secara linguistik, keunggulan IndoBERT dapat diatribusikan pada mekanisme self-attention yang memungkinkannya memahami dependensi kontekstual antar kata secara mendalam. Ulasan hotel sering kali mengandung opini implisit atau bernuansa halus, seperti "kamar agak panas" atau "pelayanan kurang cepat", yang sulit ditangkap oleh pendekatan berbasis frekuensi kata (TF-IDF) maupun embedding statis (SBERT) yang tidak diadaptasi secara spesifik. Melalui proses fine-tuning, IndoBERT mampu menginternalisasi pola linguistik khas domain perhotelan, menghasilkan representasi fitur yang jauh lebih presisi. Temuan ini sejalan dengan observasi Minaee dkk [17] yang menekankan keunggulan arsitektur Transformer dalam menangkap konteks kalimat penuh dibandingkan metode konvensional.

Kemampuan IndoBERT dalam menggeneralisasi pola pada kelas minoritas menunjukkan bahwa efektivitas model tidak hanya bergantung pada ukuran parameter, tetapi juga pada kualitas representasi yang dipelajari dari korpus bahasa Indonesia. Berbeda dengan baseline yang masih mengalami kesulitan dalam mendeteksi kelas minoritas, IndoBERT memperlihatkan ketangguhan tinggi terhadap ketidakseimbangan distribusi data. Hal ini menguatkan temuan Cahyawijaya dkk [15] ahwa model pre-trained berbasis bahasa lokal memiliki stabilitas yang lebih baik dalam menghadapi variasi label dan konteks linguistik yang khas.

Analisis visual (Gambar 3–5) memperlihatkan bahwa IndoBERT menunjukkan stabilitas kinerja antar-aspek yang tinggi (wifi, service, AC, dan kebersihan), sementara aspek bau dan general tetap menjadi tantangan karena sifat semantik yang ambigu dan distribusi datanya yang timpang. Hasil confusion matrix pada aspek bau (Gambar 6) memperlihatkan bahwa model masih cenderung memprediksi label mayoritas (netral), menyebabkan kesalahan klasifikasi pada kelas minoritas (negatif dan positif). Hal ini menegaskan bahwa meskipun teknik oversampling efektif dalam menyeimbangkan distribusi data, pendekatan tersebut belum sepenuhnya menyelesaikan kesulitan semantik pada aspek sensorik yang abstrak [25].

Selanjutnya, hasil uji signifikansi statistik menggunakan Paired T-Test dan Wilcoxon Signed-Rank Test menunjukkan bahwa perbedaan kinerja IndoBERT terhadap baseline signifikan secara statistik ($p < 0.05$), baik terhadap TF-IDF maupun SBERT. Sebaliknya, perbandingan antar-baseline tidak menunjukkan perbedaan yang signifikan ($p > 0.05$), yang mengindikasikan bahwa keunggulan IndoBERT berasal dari kapasitas arsitekturnya, bukan sekadar variasi teknik penyeimbangan data. Nilai p yang rendah (0.0019 pada uji Wilcoxon) memperkuat validitas temuan ini secara inferensial.

Sintesis dari seluruh temuan tersebut menegaskan bahwa pendekatan berbasis Transformer kontekstual, khususnya IndoBERT, menawarkan solusi yang lebih stabil, seimbang, dan reliabel untuk tugas Aspect-Based Sentiment Analysis (ABSA) berbahasa Indonesia. Penelitian ini tidak hanya membuktikan efektivitas fine-tuning dalam meningkatkan performa evaluasi, tetapi juga menegaskan urgensi penggunaan model bahasa lokal (local pre-trained models) untuk menangani nuansa linguistik bahasa Indonesia secara lebih akurat. Dengan demikian, IndoBERT dapat dianggap sebagai baseline baru yang kuat untuk penelitian ABSA di domain ulasan hotel dan potensi aplikasi lintas sektor.

IV. KESIMPULAN

Penelitian ini mengevaluasi efektivitas model IndoBERT dalam tugas Aspect-Based Sentiment Analysis (ABSA) pada ulasan hotel berbahasa Indonesia dan membandingkannya dengan dua pendekatan baseline, yaitu TF-IDF + Logistic Regression dan SBERT + Logistic Regression. Fokus utama penelitian ini terletak pada upaya mengatasi ketidakseimbangan data melalui Random Oversampling pada tingkat teks sebelum proses fine-tuning. Pendekatan ini terbukti mampu meningkatkan performa model secara signifikan, dengan capaian akurasi 0.97 dan macro F1-score 0.92.

Secara empiris, hasil penelitian menunjukkan bahwa IndoBERT memiliki stabilitas prediksi yang tinggi pada sebagian besar aspek (wifi, service, AC, dan kebersihan), sementara tantangan masih ditemukan pada aspek bau dan general yang bersifat lebih ambigu. Hasil uji signifikansi statistik menegaskan bahwa peningkatan performa IndoBERT terhadap baseline bersifat signifikan secara statistik ($p < 0.05$), sehingga keunggulan yang dicapai bukan disebabkan oleh kebetulan.

Secara keseluruhan, penelitian ini menyimpulkan bahwa IndoBERT merupakan model yang efektif dan unggul untuk tugas ABSA berbahasa Indonesia, khususnya ketika dikombinasikan dengan strategi oversampling yang tepat. Pendekatan berbasis Transformer terbukti tidak hanya meningkatkan metrik performa, tetapi juga memberikan hasil yang lebih konsisten dan valid secara inferensial dibandingkan metode klasik.

Keterbatasan penelitian ini terletak pada cakupan domain (ulasan hotel HoASA) dan penggunaan varian IndoBERT Base akibat keterbatasan sumber daya komputasi. Penelitian selanjutnya disarankan untuk memperluas eksperimen ke domain lintas sektor (cross-domain), mengeksplorasi augmentasi data berbasis parafrasa, atau menggunakan model bahasa generatif berskala besar untuk meningkatkan deteksi sentimen pada aspek dengan ambiguitas semantik tinggi.

DAFTAR PUSTAKA

[1] A. Chauhan, A. Sharma, and R. Mohana, "A Pre-Trained Model for Aspect-based Sentiment Analysis Task: using Online Social Networking," *Procedia Comput. Sci.*, vol. 233, pp. 35–44, 2024, doi: 10.1016/j.procs.2024.03.193.

[2] K. K. Yusuf, E. Ogbuju, T. Abiodun, and F. Oladipo, "A Technical Review of the State-of-the-Art Methods in Aspect-Based Sentiment Analysis," *J. Comput. Theor. Appl.*, vol. 1, no. 3, pp. 287–298, 2024, doi: 10.62411/jcta.9999.

[3] H. T. M. Le, T. A. Phan-Thi, B. T. Nguyen, and T. Q. Nguyen, "Mining online hotel reviews using big data and machine learning: An empirical study from an emerging country," *Ann. Tour. Res. Empir. Insights*, vol. 6, no. 1, p. 100170, 2025, doi: 10.1016/j.annale.2025.100170.

[4] N. D. Wulandari, M. H. Z. Nuri, and L. Kurniasari, "Customers' Satisfaction and Preferences Using Sentiment Analysis on Traveloka: The Case of Yogyakarta Special Region Hotels," *Proc. 1st UMGESHIC Int. Semin. Heal. Soc. Sci. Humanit. (UMGESHIC-ISHSSH 2020)*, vol. 585, no. April, 2021, doi: 10.2991/assehr.k.211020.058.

[5] A. N. Azhar, "2024 11th International Conference on Advanced Informatics: Concept, Theory and Application, ICAICTA 2024," *2024 11th Int. Conf. Adv. Informatics Concept, Theory Appl. ICAICTA 2024*, 2024.

[6] G. D. Aniello, M. Gaeta, and I. La, *KnowMIS - ABSA : an overview and a reference model for applications of sentiment analysis and aspect - based sentiment analysis*, vol. 55, no. 7. Springer Netherlands, 2022. doi: 10.1007/s10462-021-10134-9.

[7] D. R. I. M. Setiadi, D. Marutho, and N. A. Setiyanto, "Comprehensive Exploration of Machine and Deep Learning Classification Methods for Aspect-Based Sentiment Analysis with Latent Dirichlet Allocation Topic Modeling," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 12–22, 2024, doi: 10.62411/faith.2024-3.

[8] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, "A Survey on Aspect-Based Sentiment Analysis : Tasks , Methods , and Challenges," pp. 1–21.

[9] D. R. I. M. Setiadi, W. Warto, A. R. Muslikh, K. Nugroho, and A. N. Safriandono, "Aspect-Based Sentiment Analysis on E-commerce Reviews using BiGRU and Bi-Directional Attention Flow," *J. Comput. Theor. Appl.*, vol. 2, no. 4, pp. 470–480, 2025, doi: 10.62411/jcta.12376.

[10] H. Wan, Y. Yang, J. Du, Y. Liu, K. Qi, and J. Z. Pan, "Target-Aspect-Sentiment Joint Detection for Aspect-Based Sentiment Analysis," 2020.

[11] R. Kusumaningrum, I. Z. Nisa, R. Jayanto, R. P. Nawangsari, and A. Wibowo, "Deep learning-based application for multilevel sentiment analysis of Indonesian hotel reviews," *Heliyon*, vol. 9, no. 6, p. e17147, 2023, doi: 10.1016/j.heliyon.2023.e17147.

[12] H. Huang and A. A. Zavareh, "Sentiment Analysis in E-Commerce Platforms : A Review of Current Techniques and Future Directions," *IEEE Access*, vol. 11, no. August, pp. 90367–90382, 2023, doi: 10.1109/ACCESS.2023.3307308.

[13] S. Taj, S. M. Daudpota, A. S. Imran, and Z. Kastrati, "Aspect-based sentiment analysis for software requirements elicitation using fine-tuned Bidirectional Encoder Representations from Transformers and Explainable Artificial Intelligence," *Eng. Appl. Artif. Intell.*, vol. 151, no. March, p. 110632, 2025, doi: 10.1016/j.engappai.2025.110632.

[14] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *COLING 2020 - 28th Int. Conf. Comput. Linguist. Proc. Conf.*, pp. 757–770, 2020, doi: 10.18653/v1/2020.coling-main.66.

[15] S. Cahyawijaya *et al.*, "IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation," *EMNLP 2021 - 2021 Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 8875–8898, 2021, doi: 10.18653/v1/2021.emnlp-main.699.

[16] S. Cahyaningtyas, D. Hatta Fudholi, and A. Fathan Hidayatullah, "Deep Learning for Aspect-Based Sentiment Analysis on Indonesian Hotels Reviews," *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control*, vol. 4, no. 3, 2021, doi: 10.22219/kinetik.v6i3.1300.

[17] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning Based Text Classification: A Comprehensive Review," vol. 1, no. 1, pp. 1–43, 2021, [Online]. Available: http://arxiv.org/abs/2004.03705

[18] M. Y. Urrochman, H. Asy, and A. Ro, "Aspect-Based Sentiment Analysis of Tumpak Sewu Waterfall Tourist Reviews Using the Naive Bayes Classifier (NBC) Method," vol. 4, no. October, pp. 18–26, 2025.

[19] M. Siino, I. Tinnirello, and M. La Cascia, "Is text preprocessing still worth the time ? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers," *Inf. Syst.*, vol. 121, no. July 2023, p. 102342, 2024, doi: 10.1016/j.is.2023.102342.

[20] A. Condor, M. Litster, and Z. Pardos, "Automatic short answer grading with SBERT on out-of-sample questions," *Proc. 14th Int. Conf. Educ. Data Mining, EDM 2021*, no. Edm, pp. 345–352, 2021.

[21] L. A. Kumar and D. K. Renuka, "State-of-the-Art Natural Language Processing," *Deep Learn. Approach Nat. Lang. Process. Speech, Comput. Vis.*, pp. 49–75, 2023, doi: 10.1201/9781003348689-3.

[22] S. Ali, G. Wang, and S. Riaz, "Aspect Based Sentiment Analysis of Ridesharing Platform Reviews for Kansei Engineering," vol. 8, 2020, doi: 10.1109/ACCESS.2020.3025823.

[23] D. Marutho, Muljono, S. Rustad, and Purwanto, "Optimizing aspect-based sentiment analysis using sentence embedding transformer, bayesian search clustering, and sparse attention mechanism," *J. Open Innov. Technol. Mark. Complex.*, vol. 10, no. 1, p. 100211, 2024, doi: 10.1016/j.joitmc.2024.100211.

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

[25] P. Sundarreson and S. Kumarapathirage, "SentiGEN: Synthetic Data Generator for Sentiment Analysis," *J. Comput. Theor. Appl.*, vol. 1, no. 4, pp. 461–477, Apr. 2024, doi: 10.62411/jcta.10480.