ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

ANALISIS SENTIMEN ULASAN WISATAWAN TERHADAP DESTINASI WISATA DI SUMATERA UTARA MENGGUNAKAN INDOBERT

[1]Ommi Alfina, [2]M. Safii✉
[1]Sistem Informasi, Universitas Pembinaan Masyarakat Indonesia (UPMI), Medan, Indonesia
[2]Informatika, STIKOM Tunas Bangsa, Pematangsiantar, Indonesia
Email: m.safii@stikomtunasbangsa.ac.id

DOI: https://doi.org/10.46880/jmika.Vol10No2.pp910-918

ABSTRACT

The growth of tourist reviews on travel destinations generates textual information that can be used to understand tourists' perceptions, but sentiment analysis faces challenges such as class distribution imbalance, which can lead to bias toward the majority class. Previous studies have shown the effectiveness of IndoBERT in tourism sentiment analysis, but most have focused on achieving overall performance or have applied strategies to handle imbalance, so critical evaluation of IndoBERT's baseline behavior on highly imbalanced data is still limited. This study aims to analyze the performance of fine-tuning IndoBERT while also looking at how class imbalance affects the ability to classify tourist review sentiments. A total of 1,500 reviews were classified into positive, neutral, and negative sentiments, then evaluated using accuracy, precision, recall, F1-score, classification report, and confusion matrix. The test results showed an accuracy of 72.67%, precision of 52.80%, recall of 72.67%, and F1-score of 61.16%. Class-level evaluation showed that the model only recognized the neutral class well, with a recall of 100% and F1-score of 84.17%, while the negative and positive classes got an F1-score of 0%. This situation is related to the dominance of the neutral class at 73.07%. The contribution of this research is providing empirical evidence that accuracy can be misleading on imbalanced data and offering a diagnostic baseline for developing class imbalance strategies in tourist review sentiment analysis.

Keywords: IndoBERT, Sentiment Analysis, Traveler Reviews, Class Imbalance, Text Classification.

ABSTRAK

Pertumbuhan ulasan wisatawan pada destinasi pariwisata menghasilkan informasi tekstual yang dapat digunakan untuk memahami persepsi wisatawan, tetapi analisis sentimen menghadapi tantangan berupa ketidakseimbangan distribusi kelas yang dapat menyebabkan bias terhadap kelas mayoritas. Penelitian terdahulu telah menunjukkan efektivitas IndoBERT pada analisis sentimen pariwisata, namun sebagian besar berfokus pada pencapaian performa agregat atau telah menerapkan strategi penanganan ketidakseimbangan, sehingga evaluasi kritis terhadap perilaku baseline IndoBERT pada data yang sangat tidak seimbang masih terbatas. Penelitian ini bertujuan menganalisis kinerja fine-tuning IndoBERT sekaligus mengidentifikasi pengaruh ketidakseimbangan kelas terhadap kemampuan klasifikasi sentimen ulasan wisatawan. Sebanyak 1.500 ulasan diklasifikasikan ke dalam sentimen positif, netral, dan negatif, kemudian dievaluasi menggunakan accuracy, precision, recall, F1-score, classification report, dan confusion matrix. Hasil pengujian menunjukkan accuracy 72,67%, precision 52,80%, recall 72,67%, dan F1-score 61,16%. Namun, evaluasi per kelas menunjukkan bahwa model hanya mengenali kelas netral dengan baik, dengan recall 100% dan F1-score 84,17%, sedangkan kelas negatif dan positif memperoleh F1-score 0%. Kondisi tersebut berkaitan dengan dominasi kelas netral sebesar 73,07%. Kontribusi penelitian ini adalah memberikan bukti empiris bahwa accuracy dapat memberikan gambaran yang menyesatkan pada data tidak seimbang serta menyediakan baseline diagnostik untuk pengembangan strategi class imbalance pada analisis sentimen ulasan wisatawan.

Kata Kunci: IndoBERT, Analisis Sentimen, Ulasan Wisatawan, Class Imbalance, Klasifikasi Teks.

PENDAHULUAN

Perkembangan teknologi digital telah mengubah cara wisatawan memperoleh informasi, mengevaluasi destinasi, dan menyampaikan pengalaman setelah melakukan perjalanan. Salah satu bentuk informasi yang semakin penting adalah ulasan wisatawan pada platform digital. Ulasan tersebut tidak hanya berisi penilaian terhadap suatu destinasi, tetapi juga merepresentasikan pengalaman wisatawan terhadap berbagai aspek, seperti daya tarik destinasi, kebersihan, fasilitas, pelayanan, kenyamanan, dan kualitas lingkungan. Dengan jumlah ulasan yang terus

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

bertambah, analisis secara manual menjadi kurang efisien sehingga diperlukan pendekatan komputasional yang mampu mengekstraksi informasi dan persepsi wisatawan secara otomatis. Analisis sentimen merupakan salah satu pendekatan Natural Language Processing (NLP) yang dapat digunakan untuk mengidentifikasi kecenderungan opini dalam data tekstual (Munir, 2025; Nurwanda et al., 2024; Tampubolon et al., 2026). Dalam konteks pariwisata analisis sentimen dapat mengelompokkan ulasan ke dalam sentimen positif, netral, atau negatif sehingga pengelola destinasi dapat memperoleh gambaran yang lebih terstruktur mengenai pengalaman wisatawan. Pemanfaatan ulasan digital dengan demikian memiliki nilai strategis karena dapat mendukung evaluasi kualitas destinasi dan pengambilan keputusan berbasis data. Bahasa Indonesia pada ulasan digital dapat mengandung variasi kosakata, struktur kalimat informal, singkatan, dan konteks yang bergantung pada hubungan antarkata. Model berbasis bag-of-words atau representasi frekuensi kata memiliki keterbatasan dalam memahami konteks semantik tersebut (Putri et al., 2025; S, 2026). Perkembangan arsitektur Transformer khususnya Bidirectional Encoder Representations from Transformers (BERT), memberikan pendekatan yang lebih baik karena representasi kata dibentuk berdasarkan konteks keseluruhan kalimat (Bangkalang & Setiyawati, 2025; Gupta, 2024; Islam et al., 2026; Syah et al., 2025). Untuk bahasa Indonesia IndoBERT dikembangkan sebagai model pre-trained language model yang secara khusus dilatih menggunakan korpus berbahasa Indonesia. IndoBERT merupakan bagian dari pengembangan sumber daya NLP Indonesia melalui IndoLEM dan IndoNLP (Koto et al., 2020; Purwanto, 2026; Tandi et al., 2025).

Pemanfaatan IndoBERT untuk analisis sentimen pariwisata telah menunjukkan potensi yang cukup besar. Penelitian (Purwanto, 2026) pada ulasan Google di Kenjeran Park menggunakan 2.560 ulasan dengan distribusi 54% positif, 28% netral, dan 18% negatif. Setelah menerapkan preprocessing dan weighted loss untuk menangani ketidakseimbangan kelas, IndoBERT memperoleh akurasi 87,50% dan F1-score 0,7643. Penelitian (Singgalen, 2025) pada ulasan hotel berbahasa Indonesia juga menunjukkan bahwa IndoBERT mampu menghasilkan akurasi tinggi, tetapi dataset yang digunakan memiliki distribusi yang sangat tidak seimbang yaitu 90% positif dan masing-masing 5% untuk kelas netral dan negatif. Kondisi tersebut memperlihatkan bahwa nilai akurasi perlu ditafsirkan secara hati-hati ketika distribusi kelas tidak proporsional. Permasalahan ketidakseimbangan kelas menjadi semakin penting karena model klasifikasi dapat memperoleh akurasi yang relatif tinggi hanya dengan memprioritaskan kelas mayoritas (Benkendorf et al., 2023; Chen et al., 2024; Hellín et al., 2024). Penelitian mengenai analisis sentimen dengan BERT (Kristiyanti et al., 2024; Nasrabadi & Moattar, 2025; Nassr et al., 2025) menunjukkan imbalanced dataset memengaruhi kemampuan model dalam mengenali kelas minoritas sehingga evaluasi tidak seharusnya hanya berorientasi pada akurasi, tetapi juga mempertimbangkan precision, recall, F1-score, dan metrik agregat seperti macro F1. Penelitian terbaru pada ulasan berbahasa Indonesia bahkan menunjukkan bahwa pendekatan two-stage classification dapat meningkatkan macro F1 secara signifikan pada kondisi ketidakseimbangan ekstrem dibandingkan pendekatan single-stage dan beberapa metode penyesuaian fungsi loss (Alnahhal et al., 2026; Carvalho et al., 2025; Farhadpour et al., 2024). Permasalahan tersebut secara nyata ditemukan dalam penelitian ini. Dataset yang digunakan terdiri atas 1.500 ulasan wisatawan yang diperoleh dari ulasan wisatawan pada Google Maps terhadap wisatawan di sumatera utara, yang distribusinya didominasi oleh kelas netral sebesar 73,07%, kelas positif 18,73% dan negatif 8,20%. Setelah proses fine-tuning, IndoBERT menghasilkan akurasi pengujian sebesar 72,67%, precision 52,80%, recall 72,67%, dan F1-score 61,16%. Evaluasi per kelas menunjukkan persoalan yang lebih fundamental yaitu kelas netral memperoleh recall 100% dan F1-score 84,17%, kelas positif dan negatif memperoleh precision, recall, dan F1-score 0%. Temuan tersebut menunjukkan bahwa akurasi 72,67% pada penelitian ini terutama merefleksikan kemampuan model mengenali kelas mayoritas bukan kemampuan melakukan klasifikasi sentimen secara seimbang. Pada pengujian ulasan baru kalimat dengan indikasi positif maupun negatif masih diprediksi sebagai netral dengan tingkat confidence sekitar 70%. Permasalahan utama penelitian adalah bagaimana memperoleh nilai akurasi yang tinggi dan bagaimana membangun model analisis sentimen yang mampu mengenali seluruh kategori sentimen secara proporsional khususnya kelas positif dan negatif yang jumlah datanya lebih sedikit.

Berdasarkan kondisi tersebut research gap penelitian ini terletak pada perlunya evaluasi yang lebih kritis terhadap perilaku baseline IndoBERT pada ulasan wisatawan dengan distribusi kelas yang tidak seimbang, terutama dalam mengidentifikasi sejauh mana dominasi kelas mayoritas memengaruhi kemampuan model dalam mengenali kelas minoritas. Penelitian ini menggunakan 1.500 ulasan dengan distribusi 73,07% kelas netral, 18,73% positif, dan

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

8,20% negatif, sehingga memberikan kondisi eksperimental yang berbeda dari penelitian terdahulu dan memungkinkan analisis lebih terfokus terhadap majority-class bias.

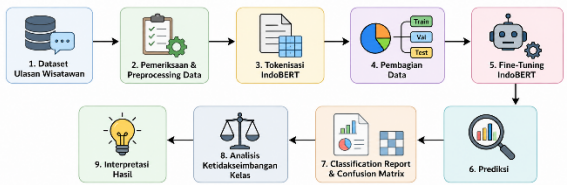
Kontribusi ilmiah penelitian ini adalah memberikan bukti empiris bahwa evaluasi performa IndoBERT pada data sentimen tidak seimbang tidak cukup hanya menggunakan akurasi sebagai indikator keberhasilan model. Hasil penelitian menunjukkan akurasi sebesar 72,67%, tetapi analisis per kelas memperlihatkan bahwa model hanya mampu mengenali kelas netral dengan baik, sedangkan kelas positif dan negatif memperoleh F1-score 0%. Dengan demikian, penelitian ini memperlihatkan secara empiris adanya kesenjangan antara performa agregat dan kemampuan klasifikasi aktual pada kelas minoritas. Temuan tersebut memperkuat pentingnya penggunaan macro-F1, classification report, dan confusion matrix sebagai bagian dari evaluasi model sentimen. Penelitian ini menghasilkan baseline diagnostik yang dapat digunakan sebagai dasar untuk penelitian selanjutnya dalam mengembangkan strategi penanganan class imbalance, seperti class weighting, data augmentation, oversampling, atau pendekatan klasifikasi bertahap. Pendekatan semacam ini juga telah menjadi arah penelitian terkini pada analisis sentimen dan klasifikasi emosi berbahasa Indonesia

TINJAUAN PUSTAKA

Penelitian terdahulu telah membuktikan bahwa IndoBERT merupakan pendekatan yang potensial untuk klasifikasi teks dan analisis sentimen berbahasa Indonesia. Penelitian pada ulasan (Purwanto, 2026) menunjukkan performa IndoBERT yang lebih baik dibandingkan LSTM, dengan akurasi 87,50% dan F1-score 0,7643 setelah menerapkan weighted loss. Penelitian pada ulasan hotel juga menunjukkan kemampuan IndoBERT dalam memperoleh akurasi tinggi, tetapi sekaligus memperlihatkan bahwa ketidakseimbangan kelas dapat menjadi persoalan penting dalam interpretasi performa. Penelitian terbaru menunjukkan bahwa strategi khusus untuk class imbalance, termasuk class weighting, SMOTE, dan klasifikasi bertahap dapat meningkatkan sensitivitas model terhadap kelas minoritas. Sebagian penelitian lebih menekankan pencapaian akurasi atau performa agregat model, sementara persoalan kegagalan model dalam mengenali kelas minoritas belum selalu menjadi fokus utama evaluasi. Performa IndoBERT pada data ulasan wisatawan dengan dominasi kelas netral yang sangat tinggi masih perlu dianalisis secara khusus. Penelitian ini mengisi celah tersebut dengan tidak hanya melaporkan performa agregat IndoBERT, tetapi juga mengevaluasi distribusi prediksi melalui classification report dan confusion matrix untuk mengidentifikasi kemampuan model pada masing-masing kelas sentimen. Kontribusi penelitian ini terletak pada evaluasi kritis terhadap kemampuan IndoBERT dalam klasifikasi sentimen wisatawan pada dataset yang tidak seimbang sekaligus mengidentifikasi keterbatasan model sebagai dasar pengembangan strategi penanganan class imbalance. Pendekatan tersebut penting agar peningkatan performa pada penelitian lanjutan tidak hanya menghasilkan akurasi yang tinggi, tetapi juga meningkatkan macro F1, recall kelas minoritas, dan keseimbangan kemampuan klasifikasi antarkelas.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan menerapkan Natural Language Processing (NLP) dan deep learning berbasis IndoBERT untuk melakukan klasifikasi sentimen terhadap ulasan wisatawan. Rancangan penelitian terdiri atas beberapa tahapan utama, yaitu pengumpulan dan persiapan data, preprocessing teks, pelabelan sentimen, pembagian dataset menjadi data pelatihan, validasi, dan pengujian, fine-tuning model IndoBERT, evaluasi performa, serta analisis kesalahan klasifikasi. Rancangan tersebut digunakan untuk mengetahui sejauh mana IndoBERT mampu memahami konteks bahasa Indonesia dalam ulasan wisatawan dan mengklasifikasikan ulasan ke dalam tiga kategori sentimen, yaitu positif, netral, dan negatif. Berdasarkan data yang tersedia, penelitian menggunakan 1.500 ulasan wisatawan dengan distribusi 1.096 ulasan netral, 281 positif, dan 123 negatif. Secara keseluruhan alur penelitian dapat dilihat pada gambar 1 berikut:



Gambar 1. Alur Penelitian

Gambar tersebut menunjukkan alur penelitian klasifikasi sentimen ulasan wisatawan menggunakan IndoBERT yang terdiri atas sembilan tahapan. Proses dimulai dari dataset ulasan wisatawan, kemudian dilakukan pemeriksaan dan preprocessing data untuk memastikan data siap digunakan. Selanjutnya, teks melalui proses tokenisasi IndoBERT dan pembagian data menjadi data pelatihan, validasi, dan pengujian.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Model kemudian dilakukan fine-tuning IndoBERT sebelum digunakan untuk menghasilkan prediksi sentimen. Hasil prediksi dievaluasi menggunakan classification report dan confusion matrix, kemudian dianalisis berdasarkan ketidakseimbangan kelas untuk mengetahui potensi bias terhadap kelas mayoritas. Tahap terakhir adalah interpretasi hasil yang digunakan untuk menarik kesimpulan mengenai kinerja dan keterbatasan model dalam mengklasifikasikan sentimen wisatawan.

Objek dan Ruang Lingkup Penelitian

Objek penelitian adalah teks ulasan wisatawan berbahasa Indonesia yang merepresentasikan pengalaman dan persepsi wisatawan terhadap destinasi wisata. Ruang lingkup penelitian dibatasi pada tugas sentiment classification tiga kelas, yaitu sentimen negatif, netral, dan positif. Variabel utama yang dianalisis adalah teks ulasan sebagai variabel masukan (input) dan kategori sentimen sebagai variabel keluaran (output). Fokus penelitian diarahkan pada identifikasi polaritas sentimen secara keseluruhan. Berdasarkan data yang tersedia terdapat ketidakseimbangan yang cukup besar antar kelas, dengan kelas netral mencapai 73,07%, sedangkan positif 18,73% dan negatif 8,20%. Kondisi ini selanjutnya diperlakukan sebagai faktor penting dalam interpretasi hasil eksperimen karena berpotensi menyebabkan model lebih dominan memprediksi kelas mayoritas.

Teknik Pengumpulan dan Persiapan Data

Data penelitian diperoleh dari ulasan wisatawan pada Google Maps terhadap wisata di Sumatera Utara yang kemudian dipersiapkan untuk proses klasifikasi. Setiap observasi terdiri atas teks ulasan dan label sentimen. Tahap persiapan meliputi pemeriksaan data, pembersihan teks, penyesuaian format, serta tokenisasi menggunakan tokenizer yang sesuai dengan model IndoBERT. Tokenisasi mengubah teks menjadi representasi token yang dapat diproses oleh model. Secara konseptual, untuk sebuah ulasan (X), proses tokenisasi dapat dinyatakan sebagai(Toraman et al., 2022):

$$X = (w_1, w_2, \dots, w_n) \quad (1)$$

Kemudian ditransformasikan menjadi rangkaian token:

$$T(X) = (t_1, t, \dots, t_n) \quad (2)$$

Dengan w_i merupakan kata atau unit teks dan t_i merupakan token yang dihasilkan oleh tokenizer IndoBERT. Karena model Transformer memiliki batas panjang input, teks selanjutnya dipotong *(truncation)* atau dilengkapi padding sehingga memiliki panjang input yang seragam.

Pelabelan dan Definisi Operasional Variabel

Variabel dependen dalam penelitian ini adalah sentimen ulasan wisatawan yang terdiri atas tiga kategori. Sentimen positif didefinisikan sebagai ulasan yang menunjukkan pengalaman, kepuasan, apresiasi, atau penilaian positif terhadap destinasi; sentimen netral merupakan ulasan yang cenderung informatif atau tidak menunjukkan polaritas emosional yang dominan; sedangkan sentimen negatif merupakan ulasan yang menunjukkan ketidakpuasan, keluhan, kekecewaan, atau penilaian buruk. Secara matematis label sentimen dapat direpresentasikan sebagai $y \in \{0,1,2\}$ dengan masing-masing nilai merepresentasikan satu kategori sentimen. Variabel independennya adalah teks ulasan wisatawan yang menjadi input bagi IndoBERT. Model menghasilkan probabilitas untuk setiap kelas menggunakan fungsi softmax(Holm et al., 2023):

$$P(y = k|x) = \frac{e^{zk}}{\sum_{j=1}^{K} e^{zj}} \quad (3)$$

dengan zk merupakan logit kelas ke- k, K=3 merupakan jumlah kelas, dan $P(y = k|x)$ merupakan probabilitas bahwa ulasan x termasuk kelas k. Label akhir ditentukan berdasarkan probabilitas terbesar:

$$y = \arg max P(y = k|X) \quad (4)$$

Fine-Tuning IndoBERT

Tahap berikutnya adalah melakukan fine-tuning IndoBERT menggunakan data pelatihan yang telah ditokenisasi. Pada tahap ini, parameter model pre-trained disesuaikan dengan karakteristik data ulasan wisatawan. Lapisan klasifikasi menghasilkan tiga logit yang masing-masing mewakili kelas negatif, netral, dan positif. Fungsi objektif yang umum digunakan dalam klasifikasi multikelas adalah cross-entropy loss yang secara matematis dirumuskan sebagai(Terven et al., 2025):

$$L = -\sum_{k=1}^{K} y_k \log(\hat{p}_k) \quad (4)$$

dengan $K = 3, y_k$ adalah label sebenarnya untuk kelas K dan $\hat{p}_k$ adalah probabilitas prediksi model untuk kelas tersebut. Untuk satu observasi, apabila label sebenarnya adalah kelas c persamaan tersebut dapat disederhanakan menjadi:

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

$$L = -\log(\hat{p}_c) \quad (5)$$

Model kemudian mengoptimalkan parameter θ dengan meminimalkan nilai loss:

$$\theta^* = \arg\min \frac{1}{N} \sum_{i=1}^{N} L_i \quad (6)$$

Teknik Analisis dan Evaluasi Model

Evaluasi model dilakukan menggunakan accuracy, precision, recall, F1-score, classification report, dan confusion matrix (Conciatori et al., 2024). Penggunaan beberapa metrik diperlukan karena distribusi kelas penelitian sangat tidak seimbang sehingga accuracy saja berpotensi memberikan gambaran performa yang menyesatkan. Accuracy mengukur proporsi prediksi yang benar terhadap keseluruhan data(Sujon et al., 2025):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

Untuk klasifikasi multikelas, accuracy dapat dinyatakan sebagai (Sujon et al., 2025):

$$Accuracy = \frac{\sum_{k=1}^{K} TP_k}{N} \quad (8)$$

dengan TP_k adalah jumlah prediksi benar pada kelas k, K adalah jumlah kelas dan N adalah jumlah seluruh data pengujian. Precision mengukur proporsi prediksi suatu kelas yang benar (Sujon et al., 2025):

$$Precission = \frac{TP}{TP+FP} \quad (9)$$

sedangkan recall mengukur kemampuan model menemukan seluruh data yang sebenarnya termasuk dalam kelas tersebut (Sujon et al., 2025):

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

Selanjutnya, F1-score merupakan rata-rata harmonis precision dan recall (Sujon et al., 2025):

$$F1 = 2x \frac{Precission\ x\ Recall}{Precission + Recall} \quad (11)$$

Karena penelitian menggunakan tiga kelas maka evaluasi juga dilakukan menggunakan macro average, yaitu rata-rata sederhana performa masing-masing kelas(Amin, 2023):

$$Macro - F1 = \frac{1}{K} \sum_{k=1}^{K} F1_k \quad (12)$$

Metrik ini sangat penting dalam penelitian karena memberikan bobot yang sama kepada kelas negatif, netral, dan positif sehingga performa pada kelas minoritas tidak tertutupi oleh dominasi kelas netral.

HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil pengolahan dan pengujian model IndoBERT dalam melakukan klasifikasi sentimen terhadap ulasan wisatawan ke dalam tiga kategori, yaitu positif, netral, dan negatif. Analisis dilakukan secara bertahap dengan memperhatikan distribusi data, kinerja model selama proses pelatihan dan pengujian, serta performa klasifikasi pada masing-masing kelas. Hasil tersebut diinterpretasikan untuk mengidentifikasi kemampuan model dalam mengenali pola sentimen, khususnya pada kondisi distribusi kelas yang tidak seimbang, serta dibandingkan dengan temuan penelitian terdahulu yang relevan.

Distribusi Data Sentimen

Hasil pengolahan dataset menunjukkan bahwa penelitian menggunakan 1.500 ulasan wisatawan yang terdiri atas 1.096 ulasan netral, 281 ulasan positif, dan 123 ulasan negatif. Secara persentase, kelas netral mendominasi dataset sebesar 73,07%, sedangkan kelas positif sebesar 18,73% dan kelas negatif sebesar 8,20%. Distribusi tersebut menunjukkan adanya ketidakseimbangan kelas yang cukup kuat, karena jumlah ulasan netral hampir empat kali lebih banyak daripada ulasan positif dan hampir sembilan kali lebih banyak daripada ulasan negatif. Kondisi ini penting karena model klasifikasi dapat terdorong untuk mempelajari pola kelas mayoritas secara lebih kuat dibandingkan kelas minoritas. Ketidakseimbangan tersebut menjadi isu utama dalam penelitian ini. Secara statistik, apabila sebuah model selalu memprediksi kelas netral, model telah memperoleh akurasi sekitar 73%, meskipun tidak mampu mengenali satu pun ulasan positif dan negatif. Dengan demikian distribusi data penelitian sejak awal menunjukkan bahwa accuracy bukan merupakan indikator yang memadai untuk menilai keberhasilan model. Penggunaan macro precision, macro recall, dan macro F1-score menjadi lebih relevan karena setiap kelas memperoleh bobot evaluasi yang sama. Hasil pengolahan dataset dapat dilihat pada gambar 2 berikut:

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

```
sentiment
Netral     1096
Positif     281
Negatif     123
Name: count, dtype: int64

Persentase:
sentiment
Netral     73.07
Positif    18.73
Negatif     8.20
Name: proportion, dtype: float64
```

Gambar 2. Distribusi Data Sentimen

Hasil Pelatihan IndoBERT

Model IndoBERT dilakukan fine-tuning untuk mempelajari karakteristik ulasan wisatawan. Hasil pelatihan menunjukkan bahwa pada epoch pertama diperoleh training loss sebesar 0,741855 dan validation loss sebesar 0,750142, dengan accuracy 73,33%, precision 53,78%, recall 73,33%, dan F1-score 62,05%. Pada epoch kedua, training loss meningkat menjadi 0,757263, sedangkan validation loss sedikit menurun menjadi 0,744590. Seluruh metrik validasi tetap berada pada nilai yang sama. Pada epoch ketiga, training loss kembali meningkat menjadi 0,765352 dan validation loss menjadi 0,753768, sementara accuracy, precision, recall, dan F1-score tetap tidak berubah. Hasil tersebut mengindikasikan bahwa penambahan epoch belum memberikan peningkatan kemampuan klasifikasi yang berarti. Perubahan loss yang relatif kecil dan metrik validasi yang stagnan menunjukkan bahwa model tidak memperoleh pembelajaran yang lebih baik terhadap kelas-kelas minoritas. Dengan kata lain, masalah utama bukan semata-mata durasi pelatihan, tetapi kemungkinan kuat berkaitan dengan distribusi data pelatihan yang tidak seimbang. Kondisi ini juga menunjukkan pentingnya melakukan evaluasi per kelas sebelum menentukan bahwa proses fine-tuning telah berhasil.

Tabel 1. Pelatihan IndoBERT

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	0.741855	0.750142	0.733333	0.537778	0.733333	0.620513
2	0.757263	0.744590	0.733333	0.537778	0.733333	0.620513
3	0.765352	0.753768	0.733333	0.537778	0.733333	0.620513

Hasil Pengujian Model

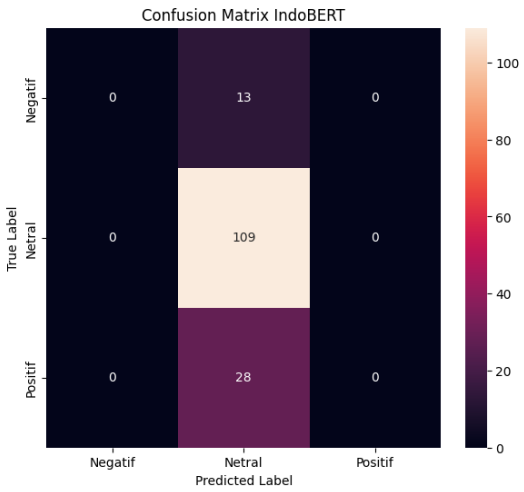
Pada data pengujian sebanyak 150 ulasan, IndoBERT menghasilkan test loss sebesar 0,7643, accuracy 72,67%, precision 52,80%, recall 72,67%, dan F1-score 61,16%. Waktu pengujian tercatat sekitar 9,93 detik dengan kecepatan 15,109 sampel per detik. Jika hanya melihat accuracy sebesar 72,67%, performa tersebut sekilas terlihat cukup baik. Namun, classification report memberikan gambaran yang berbeda. Kelas netral memperoleh precision 72,67%, recall 100%, dan F1-score 84,17%, sedangkan kelas negatif dan positif sama-sama memperoleh precision, recall, dan F1-score 0%. Macro average hanya mencapai precision 24,22%, recall 33,33%, dan F1-score 28,06%. Temuan ini merupakan hasil paling penting dalam penelitian. Accuracy 72,67% pada dasarnya sangat dekat dengan proporsi kelas netral dalam data pengujian. Dari 150 data uji, terdapat 109 data netral atau 72,67%, sehingga model yang cenderung memilih netral dapat menghasilkan accuracy tinggi tanpa benar-benar menyelesaikan tugas klasifikasi tiga kelas. Dengan demikian, nilai accuracy 72,67% tidak boleh diinterpretasikan sebagai keberhasilan klasifikasi sentimen secara menyeluruh.

```
Test Results:
test_loss: 0.7643
test_accuracy: 0.7267
test_precision: 0.5280
test_recall: 0.7267
test_f1: 0.6116
test_runtime: 9.9280
test_samples_per_second: 15.1090
test_steps_per_second: 1.9140
```

Gambar 3. Pengujian Model

Analisis Confusion Matrix dan Ketidakseimbangan Kelas

Analisis Confusion Matrix indoBERT dapat dilihat pada gambar 4 berikut:



Gambar 4. Analisis Confusion Matrix indoBERT

Berdasarkan classification report seluruh 109 ulasan yang secara aktual merupakan kelas netral berhasil dikenali sebagai netral. Sebaliknya, tidak

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

terdapat keberhasilan model dalam mengenali 13 ulasan negatif dan 28 ulasan positif sebagai kelas masing-masing. Pola tersebut memperlihatkan bahwa model mengalami bias prediksi yang sangat kuat terhadap kelas netral. Secara konseptual, kondisi tersebut dapat dijelaskan oleh fungsi optimasi klasifikasi yang berusaha meminimalkan kesalahan prediksi secara keseluruhan. Karena kelas netral memiliki jumlah observasi paling besar, kesalahan pada kelas tersebut memberikan kontribusi yang lebih besar terhadap loss dibandingkan kelas minoritas. Model kemudian memiliki kecenderungan untuk membentuk keputusan yang menguntungkan kelas mayoritas. Akibatnya, model dapat memperoleh nilai accuracy yang relatif tinggi walaupun sensitivitas terhadap kelas minoritas sangat rendah.

Eksperimen Baseline dan Improvement

Untuk menguji pengaruh ketidakseimbangan kelas terhadap kinerja IndoBERT, penelitian diperluas melalui eksperimen komparatif antara IndoBERT baseline dan IndoBERT dengan Class Weight. Kedua eksperimen menggunakan dataset, arsitektur IndoBERT, pembagian data, tokenisasi, dan parameter pelatihan yang sama sehingga perbedaan kinerja dapat dikaitkan dengan mekanisme penanganan ketidakseimbangan kelas. Berdasarkan distribusi 1.500 data dengan perhitungan bobot kelas, maka bobot yang digunakan adalah 0,4562 untuk kelas netral, 1,7794 untuk kelas positif, dan 4,0650 untuk kelas negatif. Berdasarkan hasil baseline, macro-F1 28,06% sangat penting dijadikan salah satu metrik utama karena weighted F1 61,16% terdorong oleh kelas netral yang dominan.

Perbandingan dengan Penelitian Terkait

Hasil penelitian ini berbeda dengan penelitian Purwanto pada ulasan Kenjeran Park. Penelitian tersebut menggunakan 2.560 ulasan dengan distribusi 54% positif, 28% netral, dan 18% negatif serta menerapkan weighted loss untuk mengatasi ketidakseimbangan kelas. IndoBERT pada penelitian tersebut menghasilkan accuracy 87,50% dan F1-score 0,7643. Perbedaan tersebut menunjukkan bahwa keberhasilan IndoBERT tidak hanya ditentukan oleh arsitektur model, tetapi juga oleh karakteristik dataset, distribusi kelas, preprocessing, dan strategi penanganan ketidakseimbangan data.

Temuan penelitian ini juga sejalan dengan penelitian terbaru yang secara khusus mengkaji IndoBERT pada kondisi extreme class imbalance. Tikaningsih et al. melaporkan bahwa model baseline dapat menghasilkan accuracy yang terlihat tinggi tetapi belum merepresentasikan kemampuan klasifikasi secara menyeluruh. Penelitian tersebut kemudian menggunakan two-stage classification dan memperoleh peningkatan macro F1 dari 0,599 menjadi 0,761. Hasil tersebut relevan dengan penelitian ini karena menunjukkan bahwa ketika kelas tidak seimbang, peningkatan kualitas model perlu diarahkan pada kemampuan mengenali kelas minoritas, bukan sekadar meningkatkan accuracy. Penelitian lain juga menunjukkan efektivitas strategi class weighting. Studi pada ulasan Halodoc melaporkan bahwa IndoBERT dengan class weight memperoleh macro F1 sebesar 0,8230 dan balanced accuracy sebesar 0,8274 pada kondisi kelas tidak seimbang.

Pengujian pada Ulasan Baru

Pengujian kualitatif terhadap ulasan baru dapat dilihat pada gambar 5 berikut:

```
Review     : Pemandangan Danau Toba sangat indah dan suasananya luar biasa.
Sentimen   : Netral
Confidence : 0.7119
--------------------------------------------------------------------------------
Review     : Tempatnya cukup bagus tetapi fasilitas masih perlu diperbaiki.
Sentimen   : Netral
Confidence : 0.7022
--------------------------------------------------------------------------------
Review     : Saya kecewa karena tempatnya kotor dan pelayanan sangat buruk.
Sentimen   : Netral
Confidence : 0.7009
--------------------------------------------------------------------------------
```

Gambar 5. Pengujian pada Ulasan Baru

Pengujian kualitatif terhadap ulasan baru semakin memperkuat temuan kuantitatif. Ulasan “Pemandangan Danau Toba sangat indah dan suasananya luar biasa” diklasifikasikan sebagai netral dengan confidence 0,7119. Ulasan “Tempatnya cukup bagus tetapi fasilitas masih perlu diperbaiki” juga diprediksi netral dengan confidence 0,7022. Bahkan ulasan “Saya kecewa karena tempatnya kotor dan pelayanan sangat buruk” masih dikategorikan sebagai netral dengan confidence 0,7009. Kesalahan tersebut menunjukkan bahwa persoalan model bukan hanya terjadi pada data pengujian, tetapi juga muncul ketika model diberikan data baru. Prediksi terhadap kalimat yang secara semantik memiliki kecenderungan positif atau negatif sebagai netral mengindikasikan bahwa representasi atau batas keputusan (decision boundary) model belum mampu membedakan polaritas sentimen secara memadai. Confidence sekitar 70% juga perlu ditafsirkan secara hati-hati karena confidence yang relatif tinggi tidak otomatis menunjukkan bahwa prediksi tersebut benar.

Implikasi Hasil Penelitian

Penelitian menunjukkan bahwa IndoBERT memiliki kemampuan yang baik dalam mengenali kelas mayoritas, tetapi belum mampu melakukan klasifikasi

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

sentimen tiga kelas secara seimbang. Accuracy sebesar 72,67% harus dipandang secara kritis karena diperoleh bersamaan dengan macro F1 hanya 28,06% dan F1-score 0% untuk kelas positif dan negatif. Temuan penelitian memberikan indikasi bahwa permasalahan utama terletak pada class imbalance, bukan semata-mata pada pemilihan arsitektur IndoBERT.

Hasil ini menjadi baseline yang penting untuk eksperimen berikutnya. Perbaikan model sebaiknya diarahkan pada strategi seperti class weighting, oversampling, SMOTE pada representasi yang sesuai, focal loss, atau pendekatan two-stage classification, kemudian dibandingkan menggunakan macro F1, balanced accuracy, serta recall masing-masing kelas. Pendekatan tersebut memiliki dukungan empiris dari penelitian terdahulu yang menunjukkan bahwa class weighting dan strategi klasifikasi khusus dapat meningkatkan sensitivitas terhadap kelas minoritas.

KESIMPULAN

Penelitian ini berhasil menunjukkan kemampuan dasar IndoBERT dalam memproses ulasan wisatawan, tetapi belum menghasilkan model klasifikasi sentimen yang robust. Nilai accuracy yang relatif tinggi ternyata didominasi oleh keberhasilan memprediksi kelas netral sebagai kelas mayoritas. Sebaliknya, kegagalan total dalam mengenali sentimen positif dan negatif menunjukkan adanya masalah serius pada generalisasi model terhadap kelas minoritas. Dengan demikian, kontribusi penting dari hasil penelitian ini adalah mengungkap bahwa evaluasi berbasis accuracy saja dapat menghasilkan kesimpulan yang menyesatkan pada dataset ulasan wisatawan yang tidak seimbang. Temuan ini sekaligus memberikan dasar metodologis untuk mengembangkan model IndoBERT yang lebih adil dan representatif melalui strategi penanganan class imbalance pada penelitian selanjutnya.

DAFTAR PUSTAKA

Alnahhal, M., Tabash, M. I., Safi, S. K., Al-Absy, M. S., & Mamadiyarov, Z. (2026). A Comparative Study of Imbalance-Handling Methods in Multiclass Predictive Maintenance. In *Computation* (Vol. 14, Number 4, p. 88). https://doi.org/10.3390/computation14040088

Amin, M. (2023). Confusion Matrix in Three-class Classification Problems: A Step-by-Step Tutorial. *Journal of Engineering Research*, *7*, 0. https://doi.org/10.21608/erjeng.2023.296718

Bangkalang, D., & Setiyawati, N. (2025). Perbandingan Bidirectional Encoder Representations from Transformers (BERT) Language Model pada Deteksi Emosi. *Jurnal Pendidikan Dan Teknologi Indonesia*, *5*, 2952–2960. https://doi.org/10.52436/1.jpti.988

Benkendorf, D. J., Schwartz, S. D., Cutler, D. R., & Hawkins, C. P. (2023). Correcting for the effects of class imbalance improves the performance of machine-learning based species distribution models. *Ecological Modelling*, *483*, 110414. https://doi.org/https://doi.org/10.1016/j.ecolmodel.2023.110414

Carvalho, M., Pinho, A. J., & Brás, S. (2025). Resampling approaches to handle class imbalance: a review from a data perspective. *Journal of Big Data*, *12*(1), 71. https://doi.org/10.1186/s40537-025-01119-4

Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, C. L. P. (2024). A survey on imbalanced learning: latest research, applications and future directions. *Artificial Intelligence Review*, *57*(6), 137. https://doi.org/10.1007/s10462-024-10759-6

Conciatori, M., Valletta, A., & Segalini, A. (2024). Improving the quality evaluation process of machine learning algorithms applied to landslide time series analysis. *Computers & Geosciences*, *184*, 105531. https://doi.org/https://doi.org/10.1016/j.cageo.2024.105531

Farhadpour, S., Warner, T., & Maxwell, A. (2024). Selecting and Interpreting Multiclass Loss and Accuracy Assessment Metrics for Classifications with Class Imbalance: Guidance and Best Practices. *Remote Sensing*, *16*, 533. https://doi.org/10.3390/rs16030533

Gupta, R. (2024). Bidirectional encoders to state-of-the-art: a review of BERT and its transformative impact on natural language processing. *Информатика. Экономика. Управление - Informatics. Economics. Management*, *3*, 311–320. https://doi.org/10.47813/2782-5280-2024-3-1-0311-0320

Hellín, C. J., Olmedo, A. A., Valledor, A., Gómez, J., López-Benítez, M., & Tayebi, A. (2024). Unraveling the Impact of Class Imbalance on Deep-Learning Models for Medical Image Classification. In *Applied Sciences* (Vol. 14, Number 8, p. 3419). https://doi.org/10.3390/app14083419

Holm, A. N., Wright, D., & Augenstein, I. (2023). Revisiting Softmax for Uncertainty Approximation in Text Classification. In *Information* (Vol. 14, Number 7, p. 420). https://doi.org/10.3390/info14070420

Islam, M. S., Xiangdong, L., & Ahmed, J. (2026). BERT: advancements in language understanding for different NLP tasks: challenges and future perspectives. *Journal of Electrical Systems and Information Technology*, *13*(1), 49. https://doi.org/10.1186/s43067-026-00341-1

Koto, F., Rahimi, A., Lau, J., & Baldwin, T. (2020). *IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for*

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Indonesian NLP. https://doi.org/10.18653/v1/2020.coling-main.66

Kristiyanti, D., Sanjaya, S., Tjokro, V., & Suhali, J. (2024). Dealing imbalance dataset problem in sentiment analysis of recession in Indonesia. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *13*, 2060. https://doi.org/10.11591/ijai.v13.i2.pp2060-2072

Munir, R. A. (2025). Analisis sentimen cuitan di media sosial X tentang program makan bergizi gratis dengan metode NLP. *Jurnal Informatika dan Teknik Elektro Terapan, 13*(3). https://doi.org/10.23960/jitet.v13i3.6912

Nasrabadi, H., & Moattar, M. (2025). *Sentiment Analysis of Imbalanced Dataset through Data Augmentation and Generative Annotation using DistilBERT and Low-Rank Fine-Tuning*. https://doi.org/10.21203/rs.3.rs-5879286/v1

Nassr, Z., Benabbou, F., Sael, N., & Hamim, T. (2025). Improving Sentiment Analysis Performance on Imbalanced Moroccan Dialect Datasets Using Resample and Feature Extraction Techniques. In *Information* (Vol. 16, Number 1, p. 39). https://doi.org/10.3390/info16010039

Nurwanda, N., Suarna, N., & Prihartono, W. (2024). Penerapan NLP (Natural Language Processing) Dalam Analisis Sentimen Pengguna Telegram di Playstore. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*, 1841–1846. https://doi.org/10.36040/jati.v8i2.8469

Purwanto, D. D. (2026). *Empirical Evaluation of IndoBERT and LSTM for Sentiment Analysis of Tourism Reviews: A Data-Driven Study on Kenjeran Park*. *7*(1), 463–474.

Putri, A., Sulistianingsih, N., & Rismayati, R. (2025). Pengaruh Teknik Representasi Teks Bag-of-Words dan TF-IDF terhadap Akurasi Klasifikasi Sentimen Teks Multi-Domain. *JTIM : Jurnal Teknologi Informasi Dan Multimedia*, *7*, 675–688. https://doi.org/10.35746/jtim.v7i4.756

Singgalen, Y. (2025). Performance Analysis of IndoBERT for Sentiment Classification in Indonesian Hotel Review Data. *Journal of Information System Research (JOSH)*, *6*, 976–986. https://doi.org/10.47065/josh.v6i2.6505

Sujon, K. M., Hassan, R., Choi, K., & Samad, M. A. (2025). Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models. *Journal of Big Data*, *12*(1), 268. https://doi.org/10.1186/s40537-025-01313-4

Syah, M., Puspa, A., Idhom, M., & Trimono. (2025). Perbandingan Representasi Teks Tf-Idf Dan Bert Terhadap Akurasi Cosine Similarity Dalam Penilaian Otomatis Jawaban Berbasis Teks. *Data Sciences Indonesia (DSI)*, *5*, 47–59. https://doi.org/10.47709/dsi.v5i1.6021

Tampubolon, A., Sihotang, W., Ginting, M., & Sipayung, S. (2026). Analisis Sentimen Masyarakat Pada Media Sosial X Terhadap Kepemimpinan Presiden Prabowo Subianto Tahun 2025 Menggunakan Algoritma Naïve Bayes. *Jurnal Minfo Polgan*, *14*, 3616–3625. https://doi.org/10.33395/jmp.v14i2.15837

Tandi, T. Y., Abidin, T. F., & Riza, H. (2025). Incorporation of IndoBERT and Machine Learning Features to Improve the Performance of Indonesian Textual Entailment Recognition. *Journal of Information Systems Engineering and Business Intelligence*, *11*(2 SE-Articles), 173–186. https://doi.org/10.20473/jisebi.11.2.173-186

Terven, J., Cordova-Esparza, D.-M., Romero-González, J.-A., Ramírez-Pedraza, A., & Chávez-Urbiola, E. A. (2025). A comprehensive survey of loss functions and metrics in deep learning. *Artificial Intelligence Review*, *58*(7), 195. https://doi.org/10.1007/s10462-025-11198-7

Toraman, C., Yilmaz, E., Şahinuç, F., & Ozcelik, O. (2022). *Impact of Tokenization on Language Models: An Analysis for Turkish*. https://doi.org/10.48550/arXiv.2204.08832

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.