P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

ANALISIS KARAKTERISTIK DAN KETIDAKSEIMBANGAN KELAS ULASAN TOKOPEDIA SEBAGAI DASAR FINE-TUNING INDOBERT

Alvian Totonafo Mendrofa, Nizirwan Anwar, Arief Ichwani, Raden Teddy Iswahyudi

[1,2,3,4] Program Studi Teknik Informatika Fakultas Ilmu Komputer Universitas Esa Unggul

alviantm11@student.esaunggul.ac.id, nizirwan.anwar@esaunggul.ac.id, arief.ichwani@esaunggul.ac.id, raden.teddy@esaunggul.ac.id

Abstrak

Pertumbuhan ulasan pada platform e-commerce menghasilkan data teks berbahasa Indonesia dalam jumlah besar, tetapi distribusi label yang tidak seimbang dapat menyebabkan model klasifikasi tampak memiliki performa tinggi tanpa benar-benar mampu mengenali kelas minoritas. Penelitian ini menganalisis karakteristik, kualitas, dan ketidakseimbangan kelas pada beberapa sumber ulasan Tokopedia sebagai dasar penyusunan fine-tuning IndoBERT untuk klasifikasi sentimen tiga kelas. Analisis dilakukan terhadap Tokopedia Product Reviews 2025 sebagai dataset utama serta dua sumber pembanding, yaitu Tokopedia Product Reviews 2019 dan dataset Tokopedia raw yang telah dibersihkan. Tahapan meliputi audit struktur data, pemeriksaan data kosong dan duplikasi, pemetaan rating menjadi sentimen positif, netral, dan negatif, analisis distribusi kelas, serta perhitungan majority-class baseline. Hasil menunjukkan bahwa dataset 2025 terdiri atas 65.543 ulasan dengan 97,56% positif, 1,22% netral, dan 1,22% negatif. Pada dataset raw, dari 1.346.623 entri awal diperoleh 1.072.812 ulasan unik non-kosong setelah pembersihan. Ketiga sumber menunjukkan dominasi kelas positif dengan majority baseline antara 93,23% dan 97,56%. Temuan ini menunjukkan bahwa accuracy tidak memadai sebagai satu-satunya indikator performa. Evaluasi IndoBERT perlu menekankan macro F1, precision, recall, F1-score per kelas, dan confusion matrix, khususnya untuk kelas netral dan negatif. Hasil analisis menegaskan bahwa kualitas preprocessing, konsistensi pelabelan, dan strategi evaluasi pada data tidak seimbang merupakan prasyarat penting sebelum fine-tuning IndoBERT dilakukan.

Kata kunci: analisis sentimen; IndoBERT; Tokopedia; ketidakseimbangan kelas; ulasan produk

1. PENDAHULUAN

Perdagangan elektronik mengubah cara konsumen mencari informasi, membandingkan produk, melakukan transaksi, dan menyampaikan pengalaman setelah pembelian. Salah satu jejak digital yang paling kaya adalah online customer review karena memuat penilaian terhadap kualitas produk, layanan, pengiriman, maupun pengalaman transaksi. Dalam konteks e-commerce, ulasan dapat dimanfaatkan tidak hanya oleh calon konsumen, tetapi juga sebagai sumber informasi bagi pelaku usaha untuk memahami opini pelanggan dan mengevaluasi layanan (Sari et al., 2018). Ketika jumlah review meningkat, pembacaan manual menjadi tidak efisien karena data bersifat tidak terstruktur, menggunakan kosakata informal, singkatan, variasi ejaan, serta dapat memuat lebih dari satu kecenderungan penilaian dalam satu teks.

Analisis sentimen merupakan bagian dari Natural Language Processing (NLP) yang bertujuan mengidentifikasi kecenderungan opini pada teks. Penelitian pada ulasan e-commerce Indonesia telah menggunakan pendekatan machine learning, probabilistik, deep learning, dan transformer. Stephenie et al. (2020) menganalisis ulasan produk Tokopedia menggunakan Random Forest, sedangkan Muktafin et al. (2020) menggunakan pendekatan NLP pada ulasan pembelian di Shopee. Perkembangan berikutnya menunjukkan penggunaan transformer untuk memahami konteks kalimat secara lebih baik dibandingkan representasi berbasis frekuensi kata (Bellar et al., 2024).

Untuk bahasa Indonesia, IndoBERT menjadi salah satu model pralatih yang relevan karena dirancang untuk representasi teks berbahasa Indonesia. Penerapan IndoBERT telah ditemukan pada ulasan aplikasi layanan kesehatan (Imaduddin et al., 2023), review produk Shopee (Aras et al., 2024), dan analisis sentimen berbasis aspek melalui fine-tuning IndoBERT (Widansyah et al., 2024). BERT dan turunannya juga telah digunakan pada klasifikasi ulasan pelanggan Indonesia (Nissa & Yulianti, 2023) dan pengenalan emosi pada review produk (Pramana et al., 2024). Hal tersebut menunjukkan bahwa model berbasis transformer relevan untuk tugas klasifikasi teks yang memerlukan pemahaman konteks.

Namun, keberhasilan model tidak hanya ditentukan oleh arsitektur. Distribusi label, kualitas anotasi, duplikasi, panjang review, dan ketidakseimbangan kelas dapat memengaruhi hasil secara signifikan. Dalam klasifikasi multikelas yang tidak seimbang, accuracy dapat memberikan gambaran yang terlalu optimistis karena prediksi kelas mayoritas mendominasi nilai agregat. Riyanto et al. (2023) menekankan pentingnya penggunaan beberapa metrik pada klasifikasi teks multikelas tidak seimbang. Oleh sebab itu, audit dataset perlu dilakukan sebelum fine-tuning agar hasil model dapat diinterpretasikan secara lebih adil dan dapat direproduksi.

https://doi.org/10.37817/ikraith-informatika.v10i2
P-ISSN : 2580-4316
E-ISSN : 2654-8054

Penelitian ini memusatkan perhatian pada karakteristik data ulasan Tokopedia sebagai fondasi pemodelan IndoBERT. Kontribusi penelitian terdiri atas tiga bagian. Pertama, penelitian melakukan audit komparatif terhadap tiga sumber Tokopedia untuk mengidentifikasi pola distribusi sentimen dan kualitas data. Kedua, penelitian mengukur tingkat ketidakseimbangan melalui proporsi kelas dan majority-class baseline. Ketiga, penelitian merumuskan implikasi metodologis terhadap preprocessing, pembagian data, fine-tuning, dan evaluasi IndoBERT. Fokus tersebut dipilih agar klaim performa model pada tahap berikutnya tidak terlepas dari karakter dataset yang menjadi sumber pembelajarannya.

2. METODE PENELITIAN

2.1 Desain Penelitian

Penelitian menggunakan pendekatan kuantitatif-deskriptif terhadap dataset ulasan e-commerce berbahasa Indonesia. Alur analisis meliputi identifikasi sumber data, audit struktur data, pembersihan record, pemetaan rating menjadi kelas sentimen, analisis distribusi, perhitungan majority baseline, dan penyusunan kerangka evaluasi IndoBERT. Tahap ini ditempatkan sebelum fine-tuning karena kualitas input dan distribusi target menentukan validitas eksperimen klasifikasi. Secara konseptual, pipeline penelitian mengikuti urutan dataset review, pelabelan sentimen, preprocessing, pembagian data, tokenisasi IndoBERT, fine-tuning, evaluasi model, dan implementasi aplikasi setelah model tervalidasi.

2.2 Sumber Data

Tiga sumber Tokopedia dianalisis. Tokopedia Product Reviews 2025 digunakan sebagai sumber utama karena sudah berbentuk satu review per baris dan memiliki sentiment_label. Dataset tersebut memuat 65.543 baris dan 13 kolom. Dua sumber lain digunakan sebagai pembanding karakteristik data, yaitu Tokopedia Product Reviews 2019 dengan 40.607 review dan dataset Tokopedia raw yang menyimpan daftar review pada setiap baris produk. Pada dataset raw, parsing menghasilkan 1.346.623 entri review sebelum pembersihan.

Tabel 1. Karakteristik sumber data Tokopedia

Sumber	Data Layak Analisis	Positif	Netral	Negatif
Tokopedia 2019	40.607	37.857 (93,23%)	1.825 (4,49%)	925 (2,28 %)
Tokopedia raw bersih	1.072.812	1.028.041 (95,83%)	20.416 (1,90%)	24.355 (2,27 %)
Tokopedia 2025	65.543	63.943 (97,56%)	802 (1,22%)	798 (1,22 %)

2.3 Pelabelan Sentimen

Sentimen dipetakan menjadi tiga kelas. Rating 1–2 diperlakukan sebagai sentimen negatif, rating 3 sebagai netral, dan rating 4–5 sebagai positif. Skema ini memberikan mekanisme labeling yang konsisten pada skala besar, tetapi perlu dipahami sebagai proxy sentimen berbasis rating. Ulasan dapat mengandung mixed sentiment, misalnya produk dinilai baik tetapi pengiriman dinilai lambat. Karena itu, keselarasan antara rating dan isi teks tetap menjadi salah satu sumber potensi label noise yang perlu dipertimbangkan pada analisis kesalahan.

2.4 Preprocessing dan Audit Kualitas Data

Preprocessing dibedakan menjadi pembersihan struktur data dan pembersihan teks. Pembersihan struktur mencakup pemeriksaan review kosong, duplikasi review_id, dan duplikasi teks. Pada data raw, message kosong dihapus karena tidak menyediakan input bagi model. Deduplikasi dilakukan sebelum pembagian data untuk menekan risiko data leakage. Pembersihan teks diarahkan pada operasi ringan yang relevan untuk IndoBERT, seperti lowercase bila digunakan secara konsisten, penghapusan URL dan karakter kontrol, normalisasi spasi, serta penanganan pengulangan karakter. Stopword removal dan stemming tidak diterapkan secara otomatis karena dapat menghilangkan negasi atau informasi kontekstual yang masih berguna bagi tokenizer dan model transformer (Hasanah et al., 2025).

2.5 Analisis Ketidakseimbangan dan Kerangka Evaluasi

Ketidakseimbangan kelas diukur menggunakan proporsi setiap kelas dan majority-class baseline. Majority baseline didefinisikan sebagai accuracy yang diperoleh apabila semua review diprediksi menjadi kelas mayoritas. Pada data yang sangat timpang, baseline ini penting untuk menguji apakah model benar-benar memberikan nilai tambah. Kerangka evaluasi IndoBERT dirancang menggunakan accuracy, precision, recall, dan F1-score per kelas, macro F1, weighted F1, support, serta confusion matrix. Macro F1 ditempatkan sebagai indikator utama performa

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

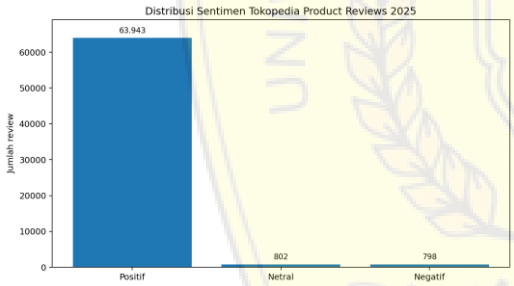
seimbang karena memberikan bobot yang sama pada setiap kelas (Riyanto et al., 2023).

3. HASIL DAN PEMBAHASAN

3.1 Karakteristik Tokopedia Product Reviews 2025

Dataset Tokopedia Product Reviews 2025 memuat 65.543 review dan 13 kolom. Kolom review_text dan sentiment_label terisi lengkap, sedangkan product_variant memiliki 38.794 nilai kosong atau sekitar 59,19% dari seluruh baris. Kekosongan product_variant tidak secara langsung menghambat klasifikasi berbasis review_text karena atribut tersebut bukan target maupun sumber teks utama. Temuan ini menunjukkan bahwa missing value perlu dinilai berdasarkan fungsi atribut, bukan hanya jumlahnya.

Distribusi kelas sangat tidak seimbang. Terdapat 63.943 review positif (97,56%), 802 netral (1,22%), dan 798 negatif (1,22%). Rasio kelas positif terhadap negatif sekitar 80,13:1 dan positif terhadap netral sekitar 79,73:1. Dengan gabungan kelas netral dan negatif hanya sekitar 2,44%, masalah utama dataset bukan kekurangan data secara keseluruhan, melainkan kelangkaan contoh minoritas dibandingkan kelas positif.

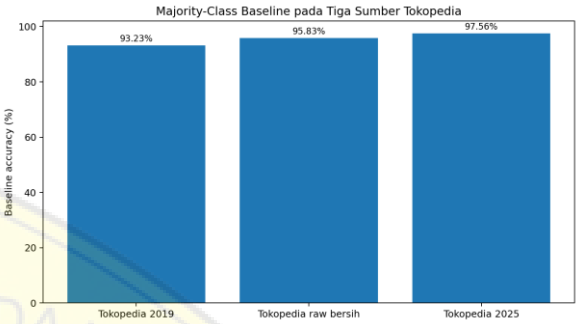


Gambar 1. Distribusi sentimen pada Tokopedia Product Reviews 2025

Dominasi kelas positif mempunyai konsekuensi langsung terhadap interpretasi performa. Jika seluruh data diprediksi positif, classifier naif sudah memperoleh accuracy sekitar 97,56%. Oleh karena itu, model dengan accuracy 97% belum dapat dianggap baik karena masih berada di bawah baseline mayoritas. Bahkan ketika accuracy melampaui baseline, kualitas model tetap harus diperiksa melalui kemampuan mengenali kelas netral dan negatif. Pada konteks analisis opini, kegagalan mendeteksi review negatif dapat menghilangkan informasi keluhan yang justru bernilai bagi pengambilan keputusan.

3.2 Perbandingan Ketidakseimbangan pada Tiga Sumber Tokopedia

Pola dominasi positif tidak hanya muncul pada dataset 2025. Tokopedia 2019 memiliki 93,23% review positif, sedangkan dataset raw yang telah dibersihkan memiliki 95,83% positif. Majority baseline ketiga sumber berada antara 93,23% dan 97,56%. Konsistensi pola tersebut menunjukkan bahwa class imbalance merupakan karakter yang berulang pada sumber data yang dianalisis, walaupun ketiga dataset tidak boleh diperlakukan sebagai satu seri waktu karena mekanisme pengumpulan dan cakupannya berbeda.



Gambar 2. Majority-class baseline pada tiga sumber Tokopedia

Tabel 2. Rasio ketimpangan kelas

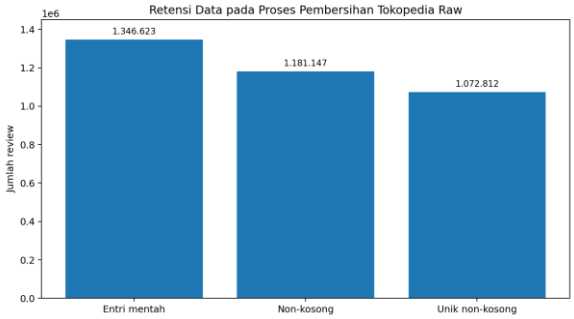
Sumber	Positif : Negatif	Positif : Netral	Positif : (Negatif +Netral)
Tokopedia 2019	40,93 : 1	20,74 : 1	13,77 : 1
Tokopedia raw bersih	42,21 : 1	50,35 : 1	22,96 : 1
Tokopedia 2025	80,13 : 1	79,73 : 1	39,96 : 1

Dataset raw bersih memiliki kelebihan dari sisi jumlah absolut kelas minoritas. Walaupun proporsi positif mencapai 95,83%, data tersebut menyediakan 20.416 review netral dan 24.355 negatif. Jumlah tersebut jauh lebih besar dibandingkan sekitar 800 contoh per kelas minoritas pada dataset 2025. Dengan demikian, ukuran total dataset tidak boleh dibaca terpisah dari jumlah support setiap kelas. Untuk fine-tuning, jumlah absolut minoritas menentukan seberapa banyak variasi bahasa keluhan dan opini netral yang dapat dipelajari model.

3.3 Kualitas Data dan Risiko Kebocoran Data

Audit terhadap dataset raw memperlihatkan bahwa kualitas struktur data memengaruhi jumlah data yang benar-benar layak digunakan. Dari 1.346.623 entri awal, sebanyak 165.476 message kosong dihapus sehingga tersisa 1.181.147 review non-kosong. Selanjutnya ditemukan 108.335 kemunculan ulang review_id pada data non-kosong. Setelah deduplikasi, diperoleh 1.072.812 review unik non-kosong atau sekitar 79,67% dari seluruh entri mentah.

https://doi.org/10.37817/ikraith-informatika.v10i2
P-ISSN : 2580-4316
E-ISSN : 2654-8054



Gambar 3. Retensi data pada proses pembersihan Tokopedia raw

Tabel 3. Audit kualitas data Tokopedia raw

Tahap	Jumlah	Proporsi	Interpretasi
Entri review mentah	1.346.623	100,00 %	Seluruh entri hasil parsing.
Pesan kosong	165.476	12,29%	Tidak menyediakan teks untuk klasifikasi.
Review non-kosong	1.181.147	87,71%	Data setelah membuang message kosong.
Duplikasi review_id	108.335	9,17% dari non-kosong	Perlu dihapus sebelum split.
Review unik non-kosong	1.072.812	79,67% dari mentah	Kandidat data setelah pembersihan.

Deduplikasi sebelum split merupakan langkah penting karena review identik yang muncul pada training dan testing dapat menyebabkan data leakage. Dalam kondisi tersebut, model mungkin tidak benar-benar melakukan generalisasi terhadap teks baru, tetapi memperoleh keuntungan dari contoh yang pernah dilihat. Dataset 2019 juga memperlihatkan 3.306 teks review yang berulang, sekitar 8,14% dari seluruh baris. Pengulangan teks tidak selalu berarti duplikasi identitas, karena komentar generik dapat ditulis oleh pengguna berbeda, tetapi tetap perlu diaudit saat merancang evaluasi.

3.4 Implikasi Pelabelan Berbasis Rating

Label berbasis rating efisien untuk membentuk target dalam jumlah besar, tetapi tidak sepenuhnya identik dengan makna teks. Pada dataset 2019 ditemukan review bintang tinggi yang tetap memuat keluhan, misalnya produk dinilai baik tetapi pengiriman lambat. Sebaliknya, review bintang rendah dapat tetap mengandung pujian pada sebagian aspek. Fenomena tersebut menunjukkan mixed sentiment dan potensi konflik antara label global dengan opini berbasis aspek. Dalam kasus seperti ini, prediksi model yang berbeda dari rating tidak selalu hanya disebabkan kelemahan model, tetapi juga dapat dipengaruhi ketidakselarasan label.

Kondisi tersebut mendukung perlunya analisis kesalahan setelah model tersedia. Contoh false positive, false negative, dan kesalahan kelas netral perlu dikelompokkan berdasarkan pola linguistik seperti teks sangat pendek, negasi, konjungsi kontras, slang, typo, pengulangan karakter, dan mixed sentiment. Pendekatan ini menghubungkan metrik kuantitatif dengan sifat bahasa review, sehingga pembahasan tidak berhenti pada angka confusion matrix.

3.5 Implikasi terhadap Fine-Tuning IndoBERT

Ketidakseimbangan kelas perlu ditangani sejak tahap pembagian data. Stratified split diperlukan agar proporsi positif, netral, dan negatif tetap terwakili pada training, validation, dan testing. Jika skema 80:10:10 digunakan pada dataset 2025, secara matematis testing hanya akan memiliki sekitar 80 review netral dan 80 review negatif. Angka tersebut bukan hasil split aktual, tetapi menggambarkan bahwa total 65 ribu review tidak otomatis berarti kelas minoritas memiliki support besar. Karena itu, support pada classification report harus selalu dilaporkan.

Pada tahap training, strategi seperti class weighting atau resampling dapat diuji sebagai pengembangan untuk mengurangi dominasi kelas positif. Namun, strategi tersebut hanya boleh diterapkan pada training set agar validation dan testing tetap merepresentasikan distribusi data yang sebenarnya. Pemilihan strategi harus dinilai melalui eksperimen terkontrol dengan konfigurasi IndoBERT yang sama. Model bahasa pralatih tetap memerlukan pencatatan parameter seperti checkpoint, epoch, batch size, learning rate, maximum sequence length, optimizer, fungsi loss, dan random seed agar eksperimen dapat direproduksi (Widansyah et al., 2024; Nur et al., 2025).

3.6 Kerangka Evaluasi IndoBERT pada Data Tidak Seimbang

Evaluasi pada penelitian ini seharusnya tidak menempatkan accuracy sebagai satu-satunya indikator. Accuracy tetap berguna untuk memberikan gambaran performa keseluruhan, tetapi harus dibandingkan dengan majority baseline. Precision per kelas mengukur ketepatan prediksi suatu kelas, recall mengukur kemampuan menemukan anggota kelas aktual, sedangkan F1-score menyeimbangkan keduanya. Macro F1 memberikan bobot yang sama pada setiap kelas, sehingga lebih sensitif terhadap kegagalan pada netral atau negatif. Weighted F1 tetap dapat dilaporkan, tetapi nilainya berpotensi didominasi oleh kelas positif karena support sangat besar.

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

Tabel 4. Kerangka evaluasi model IndoBERT

Metrik	Fungsi	Makna pada Data Tidak Seimbang
Accuracy	Proporsi seluruh prediksi yang benar.	Harus dibandingkan dengan majority baseline.
Precision per kelas	Ketepatan prediksi pada suatu kelas.	Mengukur kemurnian prediksi minoritas.
Recall per kelas	Kemampuan menemukan anggota kelas aktual.	Kritis untuk netral dan negatif.
F1-score per kelas	Keseimbangan precision dan recall.	Menunjukkan kualitas tiap kelas.
Macro F1	Rata-rata F1 dengan bobot sama.	Indikator utama performa seimbang.
Weighted F1	Rata-rata F1 berbobot support.	Dapat didominasi kelas positif.
Confusion matrix	Arah kesalahan antarkelas.	Menunjukkan kelas yang paling sering tertukar.

Temuan dari penelitian lain juga mendukung kehati-hatian tersebut. Pendekatan probabilistik maupun machine learning telah menunjukkan bahwa karakter dataset memengaruhi kinerja klasifikasi (Hayatin et al., 2022). Pada konteks review e-commerce, penelitian transformer menegaskan manfaat representasi kontekstual (Bellar et al., 2024), tetapi penggunaan model yang lebih kompleks tidak menghilangkan kebutuhan untuk mengukur error tiap kelas. IndoBERT dapat menyediakan representasi bahasa Indonesia yang lebih kontekstual, namun klaim keberhasilan tetap harus dibuktikan melalui testing pada data yang independen (Aras et al., 2024; Nissa & Yulianti, 2023).

3.7 Posisi Temuan terhadap Penelitian Terdahulu

Temuan penelitian ini dapat ditempatkan di antara dua arus utama kajian analisis sentimen e-commerce Indonesia. Arus pertama menggunakan algoritma machine learning dengan representasi fitur eksplisit. Stephenie et al. (2020), misalnya, menggunakan Random Forest pada ulasan produk Tokopedia, sedangkan Muktafin et al. (2020) menempatkan NLP sebagai dasar analisis ulasan pembelian di Shopee. Arus kedua bergerak menuju model transformer dan BERT untuk memperoleh representasi kontekstual. Aras et al. (2024) menggunakan IndoBERT pada review produk Shopee, sementara Nissa dan Yulianti (2023) menggunakan BERT pada customer review berbahasa Indonesia. Peralihan tersebut menunjukkan bahwa kompleksitas model meningkat, tetapi kualitas dataset tetap menjadi fondasi yang sama pentingnya.

Kontribusi analisis ini terletak pada penekanan terhadap kondisi sebelum training. Pada dataset Tokopedia yang diperiksa, majority baseline sudah berada pada rentang 93,23% sampai 97,56%. Nilai setinggi ini menciptakan situasi di mana model dapat terlihat unggul apabila hanya dibaca dari accuracy, padahal prediksi kelas minoritas mungkin tetap lemah. Temuan tersebut selaras dengan Riyanto et al. (2023), yang menunjukkan perlunya berbagai metrik pada klasifikasi teks multikelas tidak seimbang. Dalam konteks artikel ini, macro F1 dan recall kelas minoritas bukan sekadar metrik tambahan, melainkan bagian dari kriteria keberhasilan model.

Dari sisi data, Sutoyo et al. (2022) menunjukkan pentingnya ketersediaan dataset review produk Indonesia untuk tugas klasifikasi emosi. Analisis Tokopedia dalam penelitian ini melengkapi perspektif tersebut dengan menunjukkan bahwa besarnya jumlah review tidak otomatis menjamin distribusi label yang mendukung pembelajaran seimbang. Dataset raw yang telah dibersihkan memang menyediakan lebih dari satu juta review, tetapi 95,83% di antaranya tetap positif. Dengan demikian, volume data dan keberagaman kelas harus dinilai secara terpisah ketika menentukan kesiapan dataset untuk fine-tuning.

Dari sisi pemodelan, Widansyah et al. (2024) dan Nur et al. (2025) memperlihatkan penggunaan fine-tuning IndoBERT pada tugas sentimen Bahasa Indonesia. Artikel ini tidak membandingkan angka performa dengan kedua penelitian tersebut karena checkpoint, domain, skema label, jumlah kelas, dan distribusi data tidak identik. Sebaliknya, perbandingan dilakukan pada tingkat rancangan penelitian: fine-tuning perlu didukung split yang independen, konfigurasi yang terdokumentasi, dan evaluasi yang tidak bias terhadap kelas mayoritas. Prinsip tersebut penting agar keunggulan transformer tidak tertutup oleh kelemahan desain eksperimen.

Tabel 5. Posisi penelitian terhadap kajian terkait

Penelitian	Objek/Model	Fokus yang Relevan	Posisi terhadap Artikel Ini
Stephenie et al. (2020)	Tokopedia / Random Forest	Klasifikasi sentimen review Tokopedia.	Memberi konteks objek e-commerce Indonesia.
Aras et al. (2024)	Shopee / IndoBERT	IndoBERT pada review produk marketplace.	Memberi dasar pemilihan IndoBERT.

Penelitian	Objek/Model	Fokus yang Relevan	Posisi terhadap Artikel Ini
Riyanto et al. (2023)	Klasifikasi teks multikelas	Evaluasi pada data tidak seimbang.	Mendukung penggunaan macro F1 dan metrik per kelas.
Penelitian ini	Tokopedia / kesiapan IndoBERT	Audit kualitas data, imbalance, dan baseline.	Memperkuat fondasi data sebelum fine-tuning.

3.8 Implikasi terhadap Implementasi Aplikasi

Dalam alur penelitian yang lebih luas, model IndoBERT yang telah tervalidasi dapat diintegrasikan ke aplikasi web menggunakan Streamlit. Namun, deployment sebaiknya ditempatkan setelah model final dan tokenizer ditetapkan. Pipeline inferensi pada aplikasi harus menggunakan preprocessing, label mapping, maximum sequence length, tokenizer, dan checkpoint yang sama dengan eksperimen evaluasi. Konsistensi tersebut diperlukan agar keluaran aplikasi merepresentasikan model yang telah diuji, bukan konfigurasi berbeda yang tidak memiliki bukti performa.

Pada tahap pengujian aplikasi, fungsi minimum meliputi penerimaan input review, penanganan input kosong, prediksi sentimen, dan penyajian hasil. Pengujian fungsional perlu dibedakan dari evaluasi prediktif. Aplikasi yang berhasil menampilkan label tanpa error belum membuktikan bahwa model memiliki kualitas klasifikasi yang baik; kualitas prediksi tetap mengacu pada metrik data testing. Pemisahan antara evaluasi algoritmik dan pengujian fungsional membuat kontribusi model dan kontribusi aplikasi dapat dinilai secara tepat.

3.9 Keterbatasan Penelitian

Analisis ini memiliki beberapa keterbatasan. Pertama, tiga sumber Tokopedia memiliki struktur dan cakupan yang berbeda sehingga tidak digunakan sebagai satu seri waktu atau satu dataset gabungan. Kedua, labeling berbasis rating dapat menyederhanakan review yang memiliki sentimen campuran. Ketiga, hasil fine-tuning IndoBERT, classification report, confusion matrix, dan pengujian Streamlit belum tersedia pada dokumen sumber, sehingga artikel ini tidak menyatakan performa final model. Keterbatasan tersebut sengaja dipertahankan agar temuan artikel tidak melampaui bukti empiris yang tersedia.

4. KESIMPULAN

Analisis terhadap tiga sumber ulasan Tokopedia menunjukkan bahwa tantangan utama sebelum fine-tuning IndoBERT bukan kekurangan jumlah data, melainkan kualitas struktur data dan ketidakseimbangan kelas. Tokopedia Product Reviews 2025 terdiri atas 65.543 review dengan 97,56% kelas positif, sedangkan Tokopedia 2019 dan dataset raw bersih masing-masing memiliki 93,23% dan 95,83% kelas positif. Pada dataset raw, pembersihan message kosong dan duplikasi review_id mengurangi 1.346.623 entri awal menjadi 1.072.812 review unik non-kosong. Majority-class baseline yang mencapai 93,23%–97,56% menunjukkan bahwa accuracy tinggi tidak cukup untuk membuktikan kualitas model.

Fine-tuning IndoBERT pada data tersebut perlu didahului deduplikasi, pembagian stratified, dan dokumentasi konfigurasi yang reproduktif. Evaluasi harus menempatkan macro F1, precision, recall, F1-score per kelas, support, dan confusion matrix bersama accuracy, terutama untuk menilai kemampuan model mengenali kelas netral dan negatif. Pelabelan berbasis rating tetap dapat digunakan sebagai mekanisme operasional, tetapi mixed sentiment dan konflik rating-teks perlu diperhatikan pada analisis kesalahan. Hasil ini menyediakan dasar metodologis yang lebih kuat bagi eksperimen IndoBERT dan implementasi lanjutan pada Streamlit setelah model tervalidasi.

DAFTAR PUSTAKA

Aras, S., Yusuf, M., Ruimassa, R. Y., Wambrauw, E. A. B., & Pala'langan, E. B. (2024). Sentiment analysis on Shopee product reviews using IndoBERT. Journal of Information Systems and Informatics, 6(3), 1616–1627. https://doi.org/10.51519/journalisi.v6i3.814

Bellar, O., Baina, A., & Ballafkih, M. (2024). Sentiment analysis: Predicting product reviews for e-commerce recommendations using deep learning and transformers. Mathematics, 12(15), 2403. https://doi.org/10.3390/math12152403

Bharadwaj, L. (2023). Sentiment analysis in online product reviews: Mining customer opinions for sentiment classification. International Journal of Multidisciplinary Research, 5(5). https://doi.org/10.36948/ijfmr.2023.v05i05.6090

Hasanah, R., et al. (2025). Play Store data scrapping and preprocessing done as sentiment analysis material. Indonesian Journal of Modern Science and Technology, 1(1), 16–21. https://doi.org/10.64021/ijmst.1.1.16-21.2025

Hayatin, N., Alias, S., Hung, L. P., & Sainin, M. S. (2022). Sentiment analysis based on probabilistic classifier techniques in various Indonesian review data. Jordanian Journal of Computers and Information Technology, 8(3), 271–281. https://doi.org/10.5455/jjcit.71-1646912715

Imaduddin, H., A'la, F. Y., & Nugroho, Y. S. (2023). Sentiment analysis in Indonesian healthcare applications using IndoBERT approach. International Journal of

Advanced Computer Science and Applications, 14(8). https://doi.org/10.14569/IJACSA.2023.0140813

Muktafin, E. H., Kusrini, K., & Luthfi, E. T. (2020). Analisis sentimen pada ulasan pembelian produk di marketplace Shopee menggunakan pendekatan Natural Language Processing. Jurnal Eksplora Informatika, 10(1), 32–42. https://doi.org/10.30864/eksplora.v10i1.390

Nissa, N. K., & Yulianti, E. (2023). Multi-label text classification of Indonesian customer reviews using bidirectional encoder representations from transformers language model. International Journal of Electrical and Computer Engineering, 13(5), 5641–5652. https://doi.org/10.11591/ijece.v13i5.pp5641-5652

Nur, M. R., Wibisono, Y., & Megasari, R. (2025). Analisis sentimen dan pemodelan topik pada post tentang merek teknologi di X menggunakan fine-tuning IndoBERT dan BERTopic. Jurnal Komputer Teknologi Informasi Sistem Informasi, 4(2), 743–750. https://doi.org/10.62712/juktisi.v4i2.508

Nurhasiyah, N., Dwiyansaputra, R., Murpratiwi, S. I., & Aranta, A. (2024). Analisis sentimen pengguna platform media sosial X pada topik Pemilihan Presiden 2024 menggunakan perbandingan model monolingual dan multilingual BERT. JATI (Jurnal Mahasiswa Teknik Informatika), 9(1), 626–634. https://doi.org/10.36040/jati.v9i1.12430

Pramana, R., Jonathan, M., Yani, H. S., & Sutoyo, R. (2024). A comparison of BiLSTM, BERT, and ensemble method for emotion recognition on Indonesian product reviews. Procedia Computer Science, 245, 399–408. https://doi.org/10.1016/j.procs.2024.10.266

Pranata, J., Agustian, S., Jasril, J., & Haerani, E. (2024). Penggunaan model bahasa IndoBERT pada metode Random Forest untuk klasifikasi sentimen dengan dataset terbatas. Building of Informatics, Technology and Science, 6(3), 1668–1676. https://doi.org/10.47065/bits.v6i3.6335

Rivaldi, R. C., & Wismarini, T. D. (2024). Analisis sentimen pada ulasan produk dengan metode Natural Language Processing (NLP). Elkom: Jurnal Elektronika dan Komputer, 17(1), 120–128. https://doi.org/10.51903/elkom.v17i1.1680

Riyanto, S., Sitanggang, I. S., Djatna, T., & Atikah, T. D. (2023). Comparative analysis using various performance metrics in imbalanced data for multi-class text classification. International Journal of Advanced Computer Science and Applications, 14(6), 1082–1090. https://doi.org/10.14569/IJACSA.2023.01406116

Sari, P. K., Alamsyah, A., & Wibowo, S. (2018). Measuring e-commerce service quality from online customer review using sentiment analysis. Journal of Physics: Conference Series, 971(1), 012053. https://doi.org/10.1088/1742-6596/971/1/012053

Stephenie, Warsito, B., & Prahutama, A. (2020). Sentiment analysis on Tokopedia product online reviews using Random Forest method. E3S Web of Conferences, 202, 16006. https://doi.org/10.1051/e3sconf/202020216006

Sutoyo, R., Achmad, S., Chowanda, A., Andangsari, E. W., & Isa, S. M. (2022). PRDECT-ID: Indonesian product reviews dataset for emotions classification tasks. Data in Brief, 44, 108554. https://doi.org/10.1016/j.dib.2022.108554

Widansyah, M., Az-Zahra, F. F., & Pambudi, A. (2024). Fine-tuning model IndoBERT (Indonesian Bidirectional Encoder Representations from Transformers) untuk analisis sentimen berbasis aspek pada aplikasi M-Paspor. Joutica, 9(2), 183–195. https://doi.org/10.30736/informatika.v9i2.1310