JURNAL LOCUS: Penelitian & Pengabdian
Volume 5 No. 11 November 2026, pp 10361-10373
E-ISSN 2829-7334
DOI: 10.58344/locus.v5i11.6512

Pemanfaatan Analisis Sentimen Ulasan Pelanggan dalam Peningkatan Kualitas Layanan *E-Commerce* XYZ Menggunakan IndoBERT dan Arsitektur Hybrid

Puspita Putri Nabilah* , Ahmad Muklason
Institut Teknologi Sepuluh Nopember, Indonesia
Email : 6032242092@student.its.ac.id* , mukhlason@its.ac.id

Keywords:
Sentiment Analysis; Service Quality; E-Commerce; IndoBERT; Hybrid Architecture.

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

Copyright © 2026 by Author. Published by Riviera Publishing

Abstract

Customer reviews of e-commerce applications provide evidence about service quality, but their volume and informal language limit manual evaluation. This study analyzed sentiment in reviews of E-Commerce XYZ, compared six Indonesian-language Transformer and hybrid architectures, and developed operational service priorities. A total of 23,303 Google Play Store reviews from January 2025 to February 2026 were collected and labelled by three independent annotators (mean pairwise Cohen's kappa = 0.956). After preprocessing, 23,191 reviews were used for descriptive analysis. A deduplicated dataset of 18,609 reviews was split into training, validation, and testing sets in an 80:10:10 stratified ratio. Six models were evaluated across three random seeds. Positive sentiment accounted for 52.50% of the descriptive dataset, negative sentiment for 38.48%, and neutral sentiment for 9.02%. Delivery and Product showed the largest within-aspect negative proportions, at 58.2% and 55.0%, whereas Service contributed 76.5% of all negative reviews. IndoBERT+CNN achieved validation weighted F1 = 0.9139 and test weighted F1 = 0.8945. At a 0.9 confidence threshold, 74.4% of reviews were classified automatically at 95.81% accuracy. Combining classification with aspect-level complaints supports systematic attention to Service and targeted improvements to Delivery and Product; Price requires ongoing monitoring. These findings support decision prioritization but require operational validation before deployment.

Kata Kunci:
Analisis Sentimen;Kualitas Layanan; *E-Commerce*; IndoBERT; Arsitektur Hybrid.

Abstrak

Ulasan pelanggan pada platform *e-commerce* mengandung informasi penting untuk mengevaluasi kualitas layanan. Namun, jumlah data yang besar dan penggunaan bahasa informal menyulitkan analisis secara manual. Penelitian ini bertujuan menganalisis distribusi sentimen pelanggan *E-Commerce* XYZ, membandingkan kinerja enam arsitektur model bahasa pra-latih, serta menyusun rekomendasi peningkatan layanan. Data mencakup 23.303 ulasan Google Play Store periode Januari 2025 hingga Februari 2026, dengan pelabelan oleh tiga anotator yang menghasilkan rata-rata Cohen's Kappa sebesar 0,956. Analisis deskriptif menggunakan 23.191 ulasan setelah preprocessing, sedangkan pemodelan melibatkan 18.609 ulasan unik dengan pembagian data 80:10:10 secara stratified. Arsitektur yang diuji meliputi IndoBERT end-to-end, IndoBERT+CNN, IndoBERT+LSTM, IndoBERT+BiLSTM, IndoBERT+GAT, dan IndoRoBERTa. Setiap model dilatih menggunakan tiga seed dan dibandingkan berdasarkan rata-rata weighted F1-score validasi. Hasil menunjukkan bahwa

sentimen positif mendominasi (52,50%), diikuti negatif (38,48%) dan netral (9,02%). *Delivery* dan *Product* memiliki proporsi sentimen negatif tertinggi, masing-masing 58,2% dan 55,0%, sementara Service mencakup 76,5% dari seluruh keluhan negatif. IndoBERT+CNN menghasilkan weighted F1-score validasi sebesar 0,9139 dan pengujian sebesar 0,8945. Berdasarkan volume dan intensitas keluhan, Service memerlukan perbaikan sistemik, *Delivery* dan *Product* memerlukan perbaikan terfokus, sedangkan *Price* cukup dipantau. Dengan ambang confidence 0,9, model mampu mengklasifikasikan 74,4% data secara otomatis dengan akurasi 95,81%. Temuan ini menunjukkan potensi analisis sentimen dalam mendukung pemantauan persepsi pelanggan dan penentuan langkah perbaikan layanan.

PENDAHULUAN

Akselerasi teknologi digital dalam dua dekade terakhir telah mentransformasi lanskap perekonomian global secara fundamental serta memperkokoh peran strategis sektor perdagangan elektronik (*e-commerce*) dalam pola konsumsi harian masyarakat modern. Di Indonesia, dinamika pertumbuhan sektor perdagangan digital menunjukkan lintasan yang sangat progresif. Proyeksi riset pasar mengindikasikan bahwa jumlah pengguna aktif platform e-commerce nasional diperkirakan akan melampaui seratus juta konsumen pada tahun 2029, didorong oleh ekspansi kelas menengah perkotaan dan peningkatan penetrasi infrastruktur internet pita lebar (Statista, 2025). Pertumbuhan pesat ini mendorong akselerasi adopsi platform digital yang menawarkan kemudahan akses pemenuhan kebutuhan primer harian secara fleksibel (Gürbüz & Kotan, 2025; Ihtada et al., 2025). Secara makroekonomi, nilai transaksi digital commerce di Indonesia diproyeksikan menembus angka 313,75 miliar dolar Amerika Serikat pada tahun 2024. Sejak tahun 2018, nilai transaksi tersebut telah melonjak sebesar 212,77 miliar dolar AS dan diproyeksikan terus bertambah sebesar 390,71 miliar dolar AS sepanjang rentang waktu 2024 hingga 2030 (Statista, 2026). Tren peningkatan konsisten ini menegaskan bahwa e-commerce kini telah menjadi motor penggerak vital dalam ekosistem ekonomi digital nasional.

Eskalasi aktivitas transaksi digital yang masif secara langsung memicu ledakan volume ulasan konsumen (*online customer reviews*) pada berbagai saluran distribusi aplikasi digital, seperti Google Play Store. Ulasan publik tersebut memuat rekaman otentik mengenai pengalaman berbelanja pengguna, persepsi terhadap keandalan sistem aplikasi, efektivitas pemenuhan pesanan, hingga kecenderungan emosional pelanggan. Dalam paradigma manajemen operasi modern, data umpan balik tak terstruktur ini merepresentasikan *Voice of Customer* (VoC) yang sangat bernilai untuk mengevaluasi kinerja kualitas layanan (*service quality*) secara obyektif. Akan tetapi, membanjirnya puluhan ribu ulasan pelanggan setiap bulannya memicu tantangan analitis yang rumit. Pemrosesan dan peninjauan ulasan secara manual sudah tidak lagi memadai, memerlukan alokasi sumber daya manusia yang sangat besar, memakan waktu lama, serta rentan terhadap bias subjektivitas pengamat. Tanpa sistem analitika otomatis yang terstruktur, pihak manajemen perusahaan *e-commerce* akan kesulitan memetakan akar persoalan layanan yang sesungguhnya. Akibatnya, kebijakan perbaikan layanan berisiko bersifat reaktif, parsial, dan tidak tepat sasaran. Kondisi ini menjadi ancaman krusial bagi perusahaan *E-Commerce* XYZ yang tengah berada dalam fase krusial ekspansi pasar dan optimalisasi fitur aplikasi (Ahmadian et al., 2024).

Tantangan analisis teks di Indonesia semakin diperumit oleh karakteristik linguistik ulasan pengguna lokal yang sarat dengan bahasa percakapan informal, singkatan kata yang tidak baku, variasi slang antargenerasi, pengulangan huruf penegas emosi, serta ungkapan sarkasme kontekstual (Christian et al., 2025). Struktur bahasa nonformal ini kerap mengelabui algoritma machine learning tradisional (seperti *Naive Bayes, Support Vector Machine,* atau *Random Forest*) yang sangat bergantung pada ekstraksi fitur manual (*feature engineering*) dan tidak mampu memahami dependensi semantik antarkata (Ihtada et al., 2025; Yulianti & Nissa, 2024). Perkembangan mutakhir dalam bidang *Natural Language Processing* (NLP) menghadirkan model deep learning berbasis Transformer, seperti BERT (*Bidirectional Encoder Representations from Transformers*), yang membaca teks secara dua arah (*bidirectional context*) melalui mekanisme *self-attention* (Vaswani et al., 2023). Untuk konteks bahasa lokal, hadirnya IndoBERT menjadi terobosan penting karena model ini telah dipra-latih (*pre-trained*) menggunakan korpus teks Bahasa Indonesia berukuran masif (mencakup miliaran token dari Wikipedia, portal berita, dan korpus lokal) sehingga memiliki kepekaan semantik tinggi terhadap struktur bahasa Indonesia (Koto et al., 2020; Ihtada et al., 2025).

Meskipun IndoBERT telah terbukti unggul, klasifikasi sentimen polaritas tunggal (positif, netral, negatif) belum memberikan panduan aksi manajerial yang konkret jika tidak diintegrasikan dengan analisis aspek layanan dan penambangan frasa dominan. Selain itu, masih terdapat celah penelitian (*research gap*) empiris mengenai komparasi performa arsitektur hybrid yang mengawinkan representasi kontekstual Transformer dengan jaringan neural lanjutan seperti *Convolutional Neural Network* (CNN) untuk mengekstraksi fitur n-gram lokal, *Long Short-Term Memory* (LSTM) dan *Bidirectional* LSTM (BiLSTM) untuk menangkap ketergantungan sekuensial jangka panjang, *Graph Attention Network* (GAT) untuk memodelkan struktur relasional graf antarkata, serta varian IndoRoBERTa.

Kesenjangan penelitian terletak pada keterbatasan penerjemahan klasifikasi sentimen menjadi keputusan perbaikan layanan yang spesifik. Hasil polaritas positif, netral, dan negatif saja tidak menunjukkan aspek layanan yang menimbulkan keluhan paling besar. Di sisi lain, evaluasi arsitektur Transformer dan hybrid sering berfokus pada metrik klasifikasi, sedangkan perbandingan kinerja, variasi hasil antar-seed, pemetaan aspek, dan kemampuan pemantauan berbasis confidence belum terintegrasi dalam satu alur keputusan bagi *E-Commerce* XYZ.

Urgensi penelitian ini adalah menyediakan analisis ulasan berbahasa Indonesia yang dapat mengarahkan perbaikan layanan tanpa hanya bergantung pada pemeriksaan manual. Kebaruan yang diuji bukan penciptaan pertama model IndoBERT atau CNN, melainkan penggabungan evaluasi enam arsitektur model dengan pemetaan empat aspek layanan, ekstraksi unigram–trigram, dan matriks prioritas berdasarkan volume serta intensitas keluhan. Pendekatan ini mengaitkan evaluasi komputasional dengan kebutuhan operasional perusahaan.

Penelitian ini bertujuan mengukur distribusi sentimen pelanggan secara keseluruhan dan menurut aspek *Service, Delivery, Price,* serta *Product*; mengidentifikasi kata dan frasa dominan yang mewakili pengalaman pelanggan; membandingkan enam arsitektur IndoBERT, IndoRoBERTa, dan model hybrid menurut weighted F1-score dan kestabilan pengujian; serta menyusun prioritas perbaikan layanan berbasis temuan. Kontribusi teoretisnya adalah bukti komparatif analisis bahasa Indonesia informal, sedangkan kontribusi praktisnya ialah kerangka

pemantauan ulasan yang transparan untuk membantu manajemen menentukan tindakan berdasarkan data.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan eksperimen komparatif supervised learning untuk klasifikasi sentimen tiga kelas pada ulasan aplikasi *E-Commerce* XYZ. Data bersumber dari 23.303 ulasan *Google Play Store* periode Januari 2025–Februari 2026 yang dikumpulkan menggunakan pustaka *Python google-play-scraper* dengan pengaturan lang='id' dan country='id'. Variabel input berupa isi ulasan; keluaran klasifikasi mencakup sentimen positif, netral, atau negatif; aspek layanan dikelompokkan ke dalam *Service, Delivery, Price,* dan *Product*. Sebanyak tiga anotator independen memberi label berdasarkan pedoman yang sama; kesepakatan antar-anotator diukur dengan Cohen's Kappa berpasangan dan label akhir ditentukan melalui suara mayoritas. Tidak terdapat sampel responden yang direkrut untuk kuesioner; satuan pengamatan adalah dokumen ulasan digital.

Ulasan dibersihkan melalui pengurangan karakter berulang, case folding, penghapusan tautan dan simbol yang tidak relevan, normalisasi bahasa percakapan menggunakan kamus 1.700 pasangan kata, serta anonimisasi merek. Untuk input model Transformer, stopword dan stemming tidak dihapus agar konteks termasuk negasi tetap terbaca; penghapusan stopword terbatas pada analisis n-gram. Setelah prapemrosesan, tersedia 23.191 ulasan untuk statistik deskriptif, sedangkan 18.609 ulasan unik digunakan untuk pemodelan setelah deduplikasi guna mengurangi kebocoran data. Data pemodelan dibagi secara stratified menjadi latih, validasi, dan uji dengan perbandingan 80:10:10. Tokenisasi WordPiece digunakan untuk varian IndoBERT dan Byte-Pair Encoding untuk IndoRoBERTa, dengan panjang urutan maksimum 96 token.

Enam arsitektur yang dibandingkan adalah IndoBERT end-to-end, IndoBERT+CNN, IndoBERT+LSTM, IndoBERT+BiLSTM, IndoBERT+GAT, dan IndoRoBERTa; rincian setiap mekanisme tercantum pada Tabel 2. Model dilatih pada GPU NVIDIA Tesla T4 dengan seed 42, 123, dan 2026. Optimasi menggunakan AdamW (weight decay 0,01), warmup linear 10%, gradient clipping 1,0, dropout 0,2, serta weighted cross-entropy untuk ketidakseimbangan kelas. Pembekuan sementara lapisan awal diterapkan pada epoch pertama, pelatihan dibatasi sampai lima epoch dengan early stopping (patience dua epoch), dan learning rate encoder diuji pada 1×10^{-5}, 2×10^{-5}, serta 3×10^{-5}. Evaluasi mencakup accuracy, weighted precision, weighted recall, macro F1, weighted F1, variasi hasil antar-seed, dan cakupan otomatisasi menurut batas confidence. Analisis frekuensi n-gram dan pemetaan aspek digunakan untuk menyusun rekomendasi layanan.

Tabel 1. Deskripsi Variabel dan Parameter Penelitian

Variabel	Deskripsi Operasional	Skala Data	Peran dalam Penelitian
Review Content	Teks narasi ulasan pengguna pada Google Play Store	Tekstual / String	Variabel Input (Fitur Utama)
Sentimen	Label polaritas afektif: Positif, Netral, dan Negatif	Nominal (3 Kelas)	Variabel Target (Ground Truth)
Aspek Layanan	Kategori topik layanan: Service, Delivery, Price, Product	Nominal (4 Kategori)	Variabel Pemetaan VoC
Confidence	Nilai probabilitas Softmax maksimum kelas prediksi	Rasio [0,0 – 1,0]	Parameter Filter Automasi

Sumber: Hasil pengolahan data penelitian oleh penulis.

Tabel 2. Spesifikasi Komparatif Arsitektur Model Deep Learning

Arsitektur Model	Mekanisme Pemrosesan Representasi	Total Parameter	Parameter Tambahan
IndoBERT End-to-End	Token pooling [CLS] langsung ke dense layer	124,44 Juta	0,002 Juta
IndoBERT + CNN	Conv1D (filter 100, kernel 3,4,5) + Max Pooling	125,36 Juta	0,923 Juta
IndoBERT + LSTM	LSTM satu arah (hidden size 128) pada last token	124,90 Juta	0,460 Juta
IndoBERT + BiLSTM	Bi-directional LSTM (hidden 128x2) concatenating	125,36 Juta	0,920 Juta
IndoBERT + GAT	2-layer GAT (4 heads, window size 2) + mean pooling	124,66 Juta	0,214 Juta
IndoRoBERTa	Token <s> pooling langsung ke dense layer	124,65 Juta	0,002 Juta

Sumber: Hasil pengolahan data penelitian oleh penulis.

HASIL DAN PEMBAHASAN

Karakteristik Dataset dan Reliabilitas Anotasi

Pemeriksaan terhadap 23.303 ulasan mentah menunjukkan kelengkapan atribut yang sangat baik. Ulasan terdistribusi ke dalam empat aspek layanan: Service (19.038 ulasan / 81,70%), Price (1.779 ulasan / 7,63%), Delivery (1.611 ulasan / 6,91%), dan Product (875 ulasan / 3,75%). Dominasi aspek Service mencerminkan interaksi umum pengguna terhadap platform aplikasi, sedangkan aspek lainnya merepresentasikan pengalaman spesifik transaksi belanja.

Pengujian reliabilitas pelabelan manual oleh tiga anotator independen menghasilkan tingkat kesepakatan yang sangat tinggi. Nilai Cohen's Kappa rata-rata tercatat sebesar 0,956 (kategori almost perfect agreement). Rincian nilai kesepakatan per pasangan anotator dan per kelas sentimen disajikan pada Tabel 3.

Tabel 3. Koefisien Reliabilitas Inter-Annotator Agreement (Cohen's Kappa)

Kelas Sentimen	Anotator A & B	Anotator A & C	Anotator B & C	Rata-rata Kappa
Sentimen Negatif	0,975	0,979	0,997	0,983 (Almost Perfect)

Kelas Sentimen	Anotator A & B	Anotator A & C	Anotator B & C	Rata-rata Kappa
Sentimen Positif	0,943	0,962	0,981	0,962 (Almost Perfect)
Sentimen Netral	0,812	0,858	0,946	0,872 (Almost Perfect)
Keseluruhan Dataset	0,934	0,951	0,982	0,956 (Almost Perfect)

Sumber: Hasil pengolahan data penelitian oleh penulis

Meskipun seluruh kelas berada dalam kategori almost perfect, kelas Netral menunjukkan nilai Kappa terendah (0,872). Analisis ketidaksepakatan (disagreement) menunjukkan bahwa dari 892 ulasan (3,83%) yang mengalami silang pendapat anotator, sebanyak 69,5% terjadi antara label Netral dan Positif. Keraguan ini banyak ditemukan pada ulasan pendek tanpa konteks deskriptif mendalam seperti 'nice', 'mantul', atau 'ok', yang oleh sebagian penilai dimaknai apresiasi positif namun oleh penilai lain dianggap faktual netral.

Karakteristik panjang teks ulasan menunjukkan pola right-skewed dengan median 7 kata dan rata-rata 12,3 kata. Menariknya, ulasan bersentimen negatif memiliki median panjang 13 kata (Q1=6, Q3=25), jauh lebih panjang dibandingkan ulasan positif (median 6 kata) dan netral (median 4 kata). Konsumen cenderung menuliskan keluhan operasional secara lebih terperinci dibandingkan saat mengekspresikan kepuasan.

Dampak Preprocessing dan Reduksi Dataset

Penerapan kamus normalisasi 1.700 pasangan kata berhasil merekonstruksi 62.460 token tidak baku (22,06% dari total kata). Pasca-pembersihan, sebanyak 112 ulasan tereliminasi karena menjadi kosong (hanya berisi emoji/simbol). Selanjutnya, proses deduplikasi teks menyaring 4.546 ulasan identik guna memastikan data latih dan data uji benar-benar independen.

Deduplikasi mereduksi dataset dari 23.191 ulasan deskriptif menjadi 18.609 ulasan unik untuk pemodelan (Tabel 4). Reduksi ini menurunkan proporsi kelas positif dari 52,50% menjadi 46,78% karena banyaknya ulasan positif singkat yang identik ('bagus', 'sangat membantu'), sementara ulasan negatif meningkat proporsinya dari 38,48% menjadi 44,31%.

Tabel 4. Transformasi Distribusi Sentimen pada Setiap Tahapan Pengolahan Data

Kelas Sentimen	Data Awal Berlabel	Data Deskriptif (Pasca-Bersih)	Data Pemodelan (Deduplikasi)
Positif	12.175 (52,25%)	12.175 (52,50%)	8.706 (46,78%)
Negatif	8.926 (38,30%)	8.925 (38,48%)	8.246 (44,31%)
Netral	2.202 (9,45%)	2.091 (9,02%)	1.657 (8,90%)
TOTAL	**23.303 (100%)**	**23.191 (100%)**	**18.609 (100%)**

Sumber: Hasil pengolahan data penelitian oleh penulis.

Analisis Deskriptif Sentimen per Aspek Layanan

Untuk menjawab rumusan masalah pertama, evaluasi sebaran sentimen pada 23.191 ulasan deskriptif dianalisis secara tabulasi silang terhadap empat aspek layanan utama (Tabel 5).

Tabel 5. Tabulasi Silang Aspek Layanan dan Distribusi Sentimen Pelanggan

Aspek Layanan	Negatif	Netral	Positif	Total Ulasan	Porsi Negatif Total
Service	6.829 (36,1%)	1.547 (8,2%)	10.550 (55,7%)	18.926 (100%)	76,52 %
Delivery	938 (58,2%)	146 (9,1%)	527 (32,8%)	1.611 (100%)	10,51 %
Price	677 (38,1%)	337 (18,9%)	765 (43,0%)	1.779 (100%)	7,59 %
Product	481 (55,0%)	61 (7,0%)	333 (38,1%)	875 (100%)	5,39 %
TOTAL	**8.925 (38,5%)**	**2.091 (9,0%)**	**12.175 (52,5%)**	**23.191 (100%)**	**100,00 %**

Sumber: Hasil pengolahan data penelitian oleh penulis.

Temuan empiris pada Tabel 5 memperlihatkan kontras struktural yang tajam. Ditinjau dari intensitas keluhan (persentase ulasan negatif di dalam aspek), Delivery (58,2%) dan Product (55,0%) merupakan aspek yang paling bermasalah karena proporsi keluhannya melampaui separuh total ulasan. Namun demikian, jika ditinjau dari volume absolut, aspek Service menyumbang 76,52% (6.829 keluhan) dari seluruh ulasan negatif yang ada pada dataset. Perbedaan mendasar antara intensitas keluhan lokal dan volume keluhan sistemik inilah yang mendasari perumusan matriks prioritas perbaikan.

Analisis Kata dan Frasa Dominan sebagai Refleksi Voice of Customer (VoC)

Guna menjawab rumusan masalah kedua, dilakukan penambangan teks unigram, bigram, dan trigram untuk mengungkap narasi kontekstual yang mendasari kepuasan dan keluhan pelanggan pada masing-masing sentimen (Tabel 6).

Tabel 6. Frekuensi Frasa Dominan (Bigram & Trigram) per Kategori Sentimen

Peringkat	Sentimen Positif	Sentimen Negatif	Sentimen Netral
Top 1 Bigram	sangat membantu (1.804)	tidak bisa (1.868)	banyak promosinya (95)
Top 2 Bigram	sangat bagus (1.662)	tidak ada (1.151)	banyak promosi (73)
Top 3 Bigram	lebih mudah (758)	sangat buruk (530)	tidak bisa (63)
Top 4 Bigram	ongkos kirim (625)	tidak jelas (475)	tidak ada (52)
Top 5 Bigram	gratis ongkos (615)	lebih baik (442)	ongkos kirim (52)
Top 1 Trigram	gratis ongkos kirim (568)	1 jam sampai (98)	gratis ongkos kirim (47)
Top 2 Trigram	belanja lebih mudah (319)	gratis ongkos kirim (92)	banyak promosi menarik (16)
Top 3 Trigram	sangat membantu untuk (241)	tidak bisa masuk (88)	1 jam sampai (15)
Top 4 Trigram	belanja jadi lebih (214)	lambat sangat buruk (82)	yang malas gerak (15)
Top 5 Trigram	jadi lebih mudah (194)	malah tidak bisa (75)	tidak perlu keluar (12)

Sumber: Hasil pengolahan data penelitian oleh penulis.

Eksplorasi frasa lintas aspek mengungkap dinamika persepsi konsumen yang sangat jelas:

Aspek *Service*

Ulasan negatif didominasi keluhan teknis parah pasca-pembaruan aplikasi, seperti trigram 'lambat sangat buruk' (80), 'tidak bisa masuk' (78), dan 'tidak bisa dibuka' (68). Konsumen mengeluhkan akun yang tiba-tiba tidak terdeteksi setelah aplikasi diperbarui. Sebaliknya, apresiasi positif berpusat pada kepraktisan berbelanja dari rumah ('keluar rumah', 500; 'sangat membantu', 1.770).

Aspek *Delivery*

Frasa negatif '1 jam sampai' (52) dan 'pesan dari jam' (21) merefleksikan kegagalan perusahaan memenuhi Service Level Agreement (SLA) janji pengiriman cepat 1 jam. Konsumen mengeluhkan pesanan yang tertunda hingga berjam-jam tanpa pembaruan status kurir yang jelas ('tidak ada respons', 20; 'tidak ada konfirmasi', 13).

Aspek *Price*

Sentimen negatif terkonsentrasi pada program 'tebus murah' ('tidak ada tebus', 34; 'bisa tebus murah', 33). Konsumen merasa kecewa karena fitur tebus murah yang sebelumnya ada di versi aplikasi lama mendadak hilang atau tidak dapat diklaim saat checkout.

Aspek *Product*

Isu utama berpusat pada ketidaksesuaian fisik pemenuhan pesanan (order fulfillment), seperti stok habis sepihak ('stok', 93; 'habis', 79) dan barang yang tertinggal saat diantar kurir ('barang yang tidak', 13; 'ada barang yang', 16), yang kemudian memicu keluhan lambatnya proses refund ('pengembalian dana', 40).

Hasil Hyperparameter Tuning Encoder Learning Rate

Fine-tuning model Transformer sangat sensitif terhadap nilai learning rate encoder. Hasil pengujian pada seed acuan 42 disajikan pada Tabel 7.

Tabel 7. Hasil Hyperparameter Tuning Learning Rate Encoder (Validation Weighted F1)

Arsitektur Model	LR = 1×10^{-5}	LR = 2×10^{-5}	LR = 3×10^{-5}	LR Optimal Terpilih
IndoBERT End-to-End	0,9098	0,9147	0,9135	2×10^{-5}
IndoBERT + CNN	0,9099	0,9121	0,9085	2×10^{-5}
IndoBERT + LSTM	0,9106	0,9154	0,9113	2×10^{-5}
IndoBERT + BiLSTM	0,9109	0,9166	0,9199	3×10^{-5}
IndoBERT + GAT	0,9122	0,9122	0,9125	3×10^{-5}
IndoRoBERTa	0,9024	0,9088	0,9083	2×10^{-5}

Sumber: Hasil pengolahan data penelitian oleh penulis.

Tabel 7 menunjukkan bahwa learning rate 2×10^{-5} merupakan konfigurasi paling stabil untuk sebagian besar model. Namun, model dengan kompleksitas arsitektur lebih tinggi seperti IndoBERT+BiLSTM dan IndoBERT+GAT mencapai performa validasi puncak pada learning rate 3×10^{-5}. Nilai optimal masing-masing arsitektur inilah yang digunakan dalam pengujian multi-seed.

Evaluasi Komparatif Kinerja Enam Arsitektur Model

Guna menjawab rumusan masalah ketiga, seluruh model pada konfigurasi final dievaluasi menggunakan tiga random seed (42, 123, 2026) pada testing set independen (1.861 ulasan). Hasil komparasi disajikan pada Tabel 8.

Tabel 8. Rekapitulasi Performa Evaluasi Enam Arsitektur Model (Rata-rata ± SD)

Arsitektur Model	Validation F1-W	Test F1-Weighted	Test Accuracy	Test Macro F1	Rank Val (Uji)
IndoBERT + CNN	**0,9139 ± 0,0018**	**0,8945 ± 0,0002**	**0,8924 ± 0,0016**	**0,8190 ± 0,0048**	**1 (1)**
IndoBERT + BiLSTM	0,9137 ± 0,0054	0,8936 ± 0,0022	0,8918 ± 0,0027	0,8175 ± 0,0070	2 (2)
IndoBERT + LSTM	0,9130 ± 0,0022	0,8928 ± 0,0017	0,8911 ± 0,0020	0,8154 ± 0,0063	3 (4)
IndoBERT End-to-End	0,9129 ± 0,0018	0,8912 ± 0,0055	0,8891 ± 0,0062	0,8148 ± 0,0105	4 (5)
IndoBERT + GAT	0,9110 ± 0,0013	0,8935 ± 0,0042	0,8913 ± 0,0043	0,8177 ± 0,0096	5 (3)
IndoRoBERTa	0,9095 ± 0,0009	0,8878 ± 0,0005	0,8855 ± 0,0000	0,8074 ± 0,0025	6 (6)

Sumber: Hasil pengolahan data penelitian oleh penulis.

Berdasarkan kriteria pemilihan utama (weighted F1-score validasi tertinggi), IndoBERT + CNN terpilih sebagai model terbaik dengan nilai 0,9139. Keunggulan model ini terbukti sangat konsisten pada data uji independen dengan meraih test weighted F1-score tertinggi sebesar 0,8945 dan accuracy 0,8924. Karakteristik paling menonjol dari IndoBERT + CNN adalah stabilitas performanya yang luar biasa antarseed (standar deviasi hanya 0,0002; skor uji pada ketiga seed: 0,8947, 0,8944, 0,8944).

Mekanisme konvolusi 1D dengan ukuran kernel 3, 4, dan 5 token serta max pooling terbukti efektif menyaring sinyal frasa kunci lokal (seperti 'sangat buruk', 'tidak bisa masuk') yang sangat menentukan sentimen ulasan e-commerce yang berukuran ringkas.

Meskipun demikian, peningkatan performa model hybrid dibanding IndoBERT End-to-End terbilang marjinal (selisih F1 hanya +0,0033). Hal ini menunjukkan bahwa representasi kontekstual bawaan IndoBERT sudah sangat kaya. Di sisi lain, IndoRoBERTa mencatatkan performa terendah (F1 uji 0,8878), mengindikasikan bahwa pra-pelatihan korpus IndoBERT lebih sesuai untuk teks ulasan nonformal dibandingkan IndoRoBERTa.

Analisis Kinerja per Kelas dan Tantangan Kelas Netral

Evaluasi detail per kelas sentimen pada model terbaik IndoBERT + CNN dirangkum pada Tabel 9.

Tabel 9. Classification Report per Kelas Model Terbaik IndoBERT + CNN (Rata-rata Seed)

Kelas Sentimen	Precision	Recall	F1-Score	ROC-AUC	Average Precision
Sentimen Positif	0,9477	0,9280	0,9377	0,9810	0,9822
Sentimen Negatif	0,9098	0,9010	0,9054	0,9751	0,9667
Sentimen Netral	0,5739	0,6627	0,6139	0,9400	0,6104

Sumber: Hasil pengolahan data penelitian oleh penulis.

Tabel 9 membuktikan bahwa model memiliki keandalan luar biasa dalam mendeteksi kelas Positif (F1 = 0,9377; AUC = 0,9810) dan Negatif (F1 = 0,9054; AUC = 0,9751). Namun, performa pada kelas Netral tertinggal signifikan (F1 = 0,6139). Rendahnya performa kelas

Netral dipicu oleh dua faktor utama: jumlah sampel data latih yang minoritas (hanya 8,9%) serta batas leksikal kelas Netral yang samar, di mana banyak ulasan netral memuat kata-kata bernuansa ambigu yang membingungkan pengklasifikasi.

Analisis Error dan Kelayakan Implementasi Sistem Pemantauan

Evaluasi matriks konfusi pada data uji mengungkap bahwa dari 201 total kesalahan prediksi, sebanyak 68–70% kesalahan melibatkan kelas Netral (misalnya 56 ulasan negatif diprediksi netral, dan 43 ulasan netral diprediksi negatif).

Dari sudut pandang mitigasi risiko bisnis, kesalahan paling kritis adalah ketika keluhan pelanggan (negatif) salah diprediksi sebagai apresiasi (positif). Pada model IndoBERT + CNN, kesalahan fatal ini hanya berjumlah 26 ulasan (3,15% dari seluruh ulasan negatif aktual). Rendahnya tingkat keluhan yang lolos membuktikan keandalan model untuk pengawasan layanan.

Guna menerapkan model dalam arsitektur analitika bisnis berkelanjutan, diuji skema penyaringan berbasis ambang keyakinan (confidence threshold) probabilitas Softmax (Tabel 10).

Tabel 10. Trade-Off Cakupan Otomatisasi dan Akurasi Berdasarkan Ambang Confidence

Ambang Confidence	Cakupan Klasifikasi Otomatis	Akurasi Prediksi Otomatis	Porsi Review Review Manual
Tanpa Ambang (0,0)	100,0 %	89,20 %	0,0 %
Ambang ≥ 0,6	95,6 %	91,17 %	4,4 %
Ambang ≥ 0,7	91,8 %	92,33 %	8,2 %
Ambang ≥ 0,8	86,1 %	93,89 %	13,9 %
Ambang ≥ 0,9	**74,4 %**	**95,81 %**	**25,6 %**

Sumber: Hasil pengolahan data penelitian oleh penulis.

Penerapan ambang batas confidence ≥ 0,9 memungkinkan sistem mengklasifikasikan 74,4% ulasan masuk secara otomatis dengan akurasi sangat tinggi (95,81%). Tim Customer Service hanya perlu memverifikasi 25,6% ulasan berisiko ambigu (confidence < 0,9). Dengan rata-rata waktu inferensi hanya 4,2 milidetik per ulasan, model ini sangat layak dideploy pada sistem backend produksi e-commerce.

Matriks Prioritas dan Rekomendasi Strategis Layanan

Untuk menjawab rumusan masalah keempat, disusun Matriks Prioritas Aspek Layanan dua dimensi dengan memadukan Volume Keluhan (proporsi ulasan negatif terhadap seluruh komplain, batas rata-rata = 25%) dan Intensitas Keluhan (persentase ulasan negatif di dalam aspek, batas rata-rata dataset = 38,48%). Klasifikasi posisi aspek disajikan pada Tabel 10.

Tabel 11. Matriks Prioritas Aspek Layanan dan Arahan Strategis Operasional

Aspek Layanan	Volume Keluhan	Intensitas Keluhan	Kategori Matriks	Akar Masalah Utama	Fokus Tindakan Strategis
Service	76,52% (Tinggi)	36,08 % (Rendah)	Perbaikan Sistemik	Aplikasi lambat, gagal login pasca-update, bug sistem	Staged rollout update, regression testing, load testing
Delivery	10,51% (Rendah)	58,22 % (Tinggi)	Perbaikan Terfokus	Gagal SLA kirim 1 jam, status fiktif, CS tidak responsif	Estimasi dinamis real-time, eskalasi komplain logistik
Product	5,39 % (Rendah)	54,97 % (Tinggi)	Perbaikan Terfokus	Barang tertinggal/kurang, stok habis sepihak, refund	Sinkronisasi stok toko-aplikasi, SOP double-checking
Price	7,59% (Rendah)	38,06% (Mendekati)	Pemantauan Aktif	Fitur tebus murah hilang, transparansi syarat promo	Komunikasi promo transparan, pertahankan gratis ongkir

Sumber: Hasil pengolahan data penelitian oleh penulis.

Rekomendasi strategis dijabarkan ke dalam rencana aksi operasional berikut:

1) Rekomendasi Aspek *Service* (Perbaikan Sistemik): Menerapkan protokol peluncuran bertahap (*staged rollout*) dan pengujian regresi menyeluruh sebelum merilis pembaruan aplikasi ke publik. Manajemen wajib memprioritaskan audit otentikasi login akun pasca-pembaruan dan melakukan load testing berkala menjelang promosi belanja besar (seperti Harbolnas).
2) Rekomendasi Aspek *Delivery* (Perbaikan Terfokus): Merevisi skema SLA pengiriman '1 Jam Sampai' menjadi algoritma estimasi dinamis yang menyesuaikan jarak toko, cuaca, dan ketersediaan kurir aktif. Mengirimkan notifikasi keterlambatan otomatis dan menetapkan batas waktu respon CS maksimal 5 menit untuk isu pengiriman tertunda.
3) Rekomendasi Aspek *Product* (Perbaikan Terfokus): Memperketat sinkronisasi basis data inventaris stok fisik toko dengan katalog aplikasi secara real-time. Mewajibkan prosedur double-checking barcode sebelum serah terima kurir guna mengeliminasi insiden barang tertinggal, serta mengotomatisasi pengembalian dana (instant refund) untuk item yang gagal dipenuhi.
4) Rekomendasi Aspek *Price* (Pemantauan Aktif): Menjaga transparansi syarat dan ketentuan promosi pada antarmuka aplikasi, mengkaji ulang keberadaan program tebus murah yang sangat diminati, serta mempertahankan promosi gratis ongkos kirim yang terbukti menjadi daya tarik utama kepuasan belanja.

KESIMPULAN

Penelitian ini menunjukkan bahwa dari 23.191 ulasan yang dianalisis secara deskriptif, sentimen positif mencapai 52,50%, sentimen negatif 38,48%, dan sentimen netral 9,02%. Aspek Delivery dan Product memiliki proporsi ulasan negatif tertinggi dalam kelompoknya, masing-masing 58,2% dan 55,0%, sedangkan Service menyumbang 76,5% dari seluruh ulasan negatif. Dari enam model yang dibandingkan, IndoBERT+CNN memperoleh validation weighted F1-score sebesar 0,9139 dan test weighted F1-score sebesar 0,8945. Pada ambang

confidence 0,9, sebanyak 74,4% ulasan dapat diklasifikasikan otomatis dengan akurasi 95,81%. Temuan tersebut menunjukkan manfaat mengintegrasikan klasifikasi sentimen, pemetaan aspek layanan, dan analisis frasa untuk menentukan prioritas perbaikan, dengan Service sebagai fokus sistemik serta Delivery dan Product sebagai fokus terarah.

Implementasi operasional disarankan dilakukan secara bertahap melalui pemantauan ulasan, evaluasi kesalahan prediksi, dan pemeriksaan manusia atas ulasan dengan confidence rendah. Penelitian berikutnya perlu menguji kestabilan model pada periode data yang berbeda, perubahan kosakata, dan aplikasi e-commerce lainnya, serta mengukur dampak tindak lanjut keluhan terhadap indikator mutu layanan yang nyata. Keterbatasan penelitian meliputi ketergantungan pada ulasan aplikasi sebagai sumber data, perubahan pola bahasa dari waktu ke waktu, dan kemungkinan bias cakupan pengguna yang meninggalkan ulasan. Karena itu, keberhasilan klasifikasi tidak otomatis membuktikan peningkatan kepuasan tanpa evaluasi lapangan.

PENDANAAN

Penelitian ini dilaksanakan secara mandiri tanpa menerima hibah pembiayaan dari pihak eksternal komersial.

KONFLIK KEPENTINGAN

Seluruh penulis menyatakan bebas dari benturan kepentingan finansial maupun hubungan personal yang dapat memengaruhi objektivitas ilmiah riset ini.

KONTRIBUSI PENULIS

Puspita Putri Nabilah merancang metodologi, scraping data, pemodelan deep learning, dan penyusunan draf naskah. Ahmad Muklason melakukan supervisi penelitian, validasi formulasi statistik, dan penelaahan kritis naskah akhir.

PERNYATAAN PENGGUNAAN KECERDASAN ARTIFISIAL GENERATIF

Penulis menyatakan bahwa kecerdasan artifisial (AI) generatif digunakan dalam penyusunan manuskrip ini. Penulis mengakui penggunaan AI generatif untuk membantu meningkatkan kejelasan bahasa dan memperbaiki struktur tata bahasa manuskrip. Penggunaan tersebut dilakukan dengan mematuhi pedoman etika yang berlaku mengenai penggunaan dan pengungkapan AI generatif dalam penelitian ilmiah. Setiap penulis telah memberikan kontribusi intelektual yang substansial terhadap karya ini. Seluruh isi manuskrip telah ditelaah secara cermat untuk memastikan keakuratannya, dan penulis bertanggung jawab sepenuhnya atas integritas serta isi akhir manuskrip.

REFERENSI

Ahmadian, H., Abidin, T. F., Riza, H., & Muchtar, K. (2024). Hybrid models for emotion classification and sentiment analysis in Indonesian language. *Applied Computational Intelligence and Soft Computing*, *2024*(1), 2826773.

Christian, W., Adamlu, D., Yu, A., & Suhartono, D. (2025). Leveraging IndoBERT and DistilBERT for Indonesian emotion classification in e-commerce reviews. Procedia Computer Science, 269, 321–330.

Dzakwan Bintang Lazuardi, M., & Noor Fatyanosa, T. (2025). Klasifikasi Emosi Multikelas Berbasis Teks Bahasa Indonesia Menggunakan IndoRoBERTa. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 9(9), 1205–1214.

Farizi, A. F. Al, & Sibaroni, Y. (2025). Implementation of BiLSTM and IndoBERT for Sentiment Analysis of TikTok Reviews. JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), 10(1), 96–106.

Gürbüz, M., & Kotan, M. (2025). Multi-Category E-Commerce Insights via Social Media Analysis using Machine Learning and BERT. Acta Infologica, 2025(1), 1–18.

Ihtada, F. K., Alfianita, R., & Aziz, O. Q. (2025). Aspect-based Multilabel Classification of E-commerce Reviews Using Fine-tuned IndoBERT. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 10(1), 45–56.

Imron, S., Setiawan, E. I., Santoso, J., & Purnomo, M. H. (2023). Aspect Based Sentiment Analysis Marketplace Product Reviews Using BERT, LSTM, and CNN. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 7(3), 586–591.

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. Proceedings of the 28th International Conference on Computational Linguistics, 757–770.

Mandhasiya, D. G., Murfi, H., & Bustamam, A. (2024). The hybrid of BERT and deep learning models for Indonesian sentiment analysis. Indonesian Journal of Electrical Engineering and Computer Science, 33(1), 591–602.

Murfi, H., Theresia Gowandi, S., Ardaneswari, G., & Nurrohmah, S. (2024). BERT-based combination of convolutional and recurrent neural network for Indonesian sentiment analysis. Applied Soft Computing, 151, 111112.

Nur, M. A., Umar, N., Feng, Z., & Gani, H. (2025). Evaluation of IndoBERT and RoBERTa: Performance of Indonesian Language Transformer Models in Sentiment Classification. JIKO (Jurnal Informatika dan Komputer), 8(2), 121–127.

Olusegun, J., Chris, E., Favor, B., & Ramo, T. (2024). Aspect-Based Sentiment Analysis of Customer Reviews in E-Commerce. International Journal of Advanced Computer Science, 14(6), 112–120.

Statista. (2025). Retail Trade in Indonesia - Statistics & Facts. Statista Research Department.

Statista. (2026). Digital Commerce Transaction Value in Indonesia From 2018 to 2030. Statista Market Forecast.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2023). Attention Is All You Need. Advances in Neural Information Processing Systems, 30, 1–11.

Xin, X., Wumaier, A., Kadeer, Z., & He, J. (2023). SSEMGAT: Syntactic and Semantic Enhanced Multi-Layer Graph Attention Network for Aspect-Level Sentiment Analysis. Applied Sciences, 13(8), 5085.

Yulianti, E., & Nissa, N. K. (2024). ABSA of Indonesian customer reviews using IndoBERT: single-sentence and sentence-pair classification approaches. Bulletin of Electrical Engineering and Informatics, 13(5), 3579–3589.