

Pendekatan Hybrid Transformer-BiLSTM Berbasis Multi-Task Learning untuk Klasifikasi dan Prioritisasi Pengaduan Pelanggan

A Multi-Task Learning-Based Hybrid Transformer-BiLSTM Approach for Customer Complaint Classification and Prioritization

Reifal Hafizh Satria*, Arry Maulana Syarif

Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Dian Nuswantoro, Semarang, Indonesia

Email: ^{1,*}111202214473@mhs.dinus.ac.id, ²arry.maulana@dsn.dinus.ac.id

Email Penulis Korespondensi: 111202214473@mhs.dinus.ac.id

Received: 01/07/2026, Revised: 30/07/2026, Accepted: 29/09/2026, Published: 29/09/2026

Abstrak—Perkembangan platform e-commerce menghasilkan volume ulasan pengguna yang sangat besar sehingga berpotensi dimanfaatkan untuk memahami permasalahan pelanggan. Namun, karakteristik bahasa yang tidak formal, banyaknya variasi keluhan, serta kebutuhan untuk mengidentifikasi jenis keluhan sekaligus tingkat prioritas penanganannya menjadikan proses klasifikasi secara otomatis lebih kompleks. Oleh karena itu, penelitian ini mengembangkan model Hybrid Transformer–BiLSTM berbasis Multi-Task Learning (MTL) agar klasifikasi kategori dan prioritas dapat dipelajari secara simultan melalui representasi fitur bersama (shared representation). Dataset diperoleh melalui web scraping Google Play Store yang menghasilkan 100.000 ulasan mentah, kemudian melalui proses preprocessing sehingga diperoleh 42.929 ulasan untuk pemodelan. Pelabelan dilakukan menggunakan pendekatan pseudo labeling berbasis aturan melalui identifikasi kata kunci, scoring berbobot, dan resolusi ambiguitas. Model memanfaatkan IndoBERT sebagai transformer encoder yang dikombinasikan dengan BiLSTM untuk menangkap representasi kontekstual dan hubungan sekuensial teks. Hasil evaluasi menunjukkan akurasi klasifikasi kategori sebesar 96,17%, dengan weighted F1-score 0,9617 dan macro F1-score 0,9353. Pada klasifikasi prioritas, model memperoleh akurasi 80,04%, weighted F1-score 0,7437, dan macro F1-score 0,4236. Perbedaan performa tersebut menunjukkan bahwa klasifikasi prioritas masih dipengaruhi oleh ketidakseimbangan distribusi data, sehingga model cenderung lebih baik dalam mengenali kelas mayoritas dibandingkan kelas minoritas. Selain itu, kualitas label yang dihasilkan juga bergantung pada aturan dalam proses pseudo labeling. Meskipun demikian, hasil penelitian menunjukkan bahwa pendekatan yang diusulkan memiliki potensi sebagai dasar pengembangan sistem otomatis untuk klasifikasi dan prioritas keluhan pelanggan pada platform e-commerce. Kontribusi penelitian ini terletak pada penerapan arsitektur Hybrid IndoBERT–BiLSTM berbasis Multi-Task Learning untuk klasifikasi kategori dan prioritas keluhan pelanggan secara simultan serta penyediaan dataset keluhan Tokopedia berbahasa Indonesia yang telah melalui proses pseudo labeling sebagai referensi bagi penelitian selanjutnya.

Kata Kunci: Hybrid Transformer-BiLSTM; IndoBERT; Ulasan E-Commerce; Klasifikasi; Prioritisasi

Abstract—The rapid growth of e-commerce platforms has generated a massive volume of user reviews that can be utilized to understand customer issues. However, the informal nature of user-generated text, the diversity of complaint types, and the need to identify both complaint categories and their handling priorities make automatic complaint classification a challenging task. Therefore, this study proposes a Hybrid Transformer–BiLSTM model based on Multi-Task Learning (MTL) to simultaneously classify complaint categories and priority levels through a shared feature representation. The dataset was collected by web scraping Google Play Store reviews of the Tokopedia application, resulting in 100,000 raw reviews. After the preprocessing stage, 42,929 reviews were retained for model development. The labeling process employed a rule-based pseudo-labeling approach involving keyword identification, weighted scoring, and ambiguity resolution to generate category and priority labels. The proposed model integrates IndoBERT as the transformer encoder with BiLSTM to capture contextual representations and sequential dependencies in Indonesian text. Experimental results show that the proposed model achieved 96.17% accuracy, with a weighted F1-score of 0.9617 and a macro F1-score of 0.9353 for complaint category classification. For complaint priority classification, the model achieved 80.04% accuracy, with a weighted F1-score of 0.7437 and a macro F1-score of 0.4236. The lower performance on the priority classification task indicates that the model is still affected by class imbalance, resulting in better recognition of majority classes than minority classes. In addition, the quality of the generated labels depends on the predefined rules used in the pseudo-labeling process. Nevertheless, the proposed approach demonstrates strong potential as a foundation for developing an automated complaint classification and prioritization system for e-commerce customer service. The contribution of this study lies in the implementation of a Hybrid IndoBERT–BiLSTM architecture based on Multi-Task Learning for the simultaneous classification of customer complaint categories and priorities, as well as the provision of an Indonesian Tokopedia complaint dataset that has undergone a pseudo-labeling process as a reference for future research.

Keywords: Hybrid Transformer-BiLSTM; IndoBERT; E-Commerce Reviews; Classification; Prioritization

1. PENDAHULUAN

Pertumbuhan pesat industri e-commerce di Indonesia telah mengubah perilaku konsumen dalam berbelanja, didorong oleh penetrasi internet yang tinggi dan meningkatnya adopsi platform marketplace lokal seperti Tokopedia [1]. Indonesia tercatat sebagai salah satu pasar e-commerce terbesar di kawasan, dengan proporsi pengguna yang sangat tinggi di antara pengguna internet dan lonjakan transaksi selama masa pandemi, yang juga diikuti meningkatnya keluhan layanan e-commerce [2]. E-commerce berkontribusi signifikan terhadap pertumbuhan ekonomi dan kesejahteraan masyarakat, sehingga kualitas pengalaman dan layanan digital menjadi isu yang semakin strategis bagi pelaku industri [3]. Tokopedia, sebagai salah satu platform terbesar, melayani jutaan transaksi dan interaksi

pelanggan, sehingga pengelolaan keluhan menjadi faktor kritis dalam menjaga kepuasan, loyalitas, dan reputasi Perusahaan [1].

Peningkatan volume transaksi berbanding lurus dengan bertambahnya keluhan pelanggan terkait pengiriman, kualitas produk, fitur aplikasi, promosi, dan layanan purna jual. Keluhan yang tidak tertangani dengan baik dapat menurunkan kepuasan, meningkatkan risiko churn, dan memicu keluhan hukum maupun publik yang merugikan reputasi Perusahaan [1], [4]. Studi pada sektor jasa dan e-commerce menunjukkan bahwa kecepatan dan kualitas respons keluhan berpengaruh kuat terhadap persepsi kualitas layanan, niat pembelian ulang, dan electronic word of mouth [5], [6], [7]. Kemampuan merespons keluhan secara cepat dan tepat menjadi pembeda penting yang dapat memperkuat kepercayaan dan komitmen pelanggan jangka panjang [5], [7], [8], [9]. Secara praktis, klasifikasi keluhan secara manual memerlukan waktu dan sumber daya manusia yang besar, sehingga sering menimbulkan penundaan penanganan kasus berprioritas tinggi. Sistem klasifikasi keluhan otomatis berbasis teks menjadi salah satu pendekatan yang dapat digunakan untuk membantu proses identifikasi dan pengelompokan pengaduan secara lebih efisien [10], [11]. Aditama dan Wicaksono mengembangkan model klasifikasi keluhan dan non-keluhan pada interaksi media sosial e-commerce Indonesia menggunakan beberapa algoritma machine learning, dengan model berbasis BERT mencapai F1-score tertinggi sebesar 98,3% pada 10.600 data [12]. Meski efektif, penelitian tersebut masih berhenti pada tahap deteksi keluhan/non-keluhan dan belum mengklasifikasikan jenis keluhan maupun tingkat prioritas penanganannya. Kerangka kerja analitik untuk pelabelan dan prioritasasi keluhan berbasis deep learning dan transformer terbukti mampu mengidentifikasi keluhan urgensi tinggi [13]. Pemodelan prioritas keluhan dengan kombinasi teknik NLP dan deep learning (misalnya BERT, CNN, BiLSTM) dapat meningkatkan alokasi sumber daya layanan pelanggan [14], [15].

Pendekatan deep learning telah menjadi state-of-the-art dalam klasifikasi teks [16]. Model berbasis transformer seperti BERT dan turunannya menunjukkan kinerja unggul pada beragam tugas NLP, termasuk klasifikasi multi-label, analisis emosi, serta pemrosesan dokumen kebijakan dan layanan public [17], [18], [19]. Pada model penambangan teks berbahasa Indonesia, model berbasis transformer seperti IndoBERT telah menunjukkan kinerja yang lebih unggul dibandingkan pendekatan non-transformer dalam analisis teks, khususnya dalam memahami konteks bahasa secara lebih mendalam [20], [21]. Zevana dan Riana mengombinasikan IndoBERT fine-tuned dengan CNN dan BiLSTM untuk mengklasifikasikan opini layanan pengiriman J&T Express dari Twitter, dengan arsitektur CNN mencapai akurasi tertinggi sebesar 83% [22]. Penelitian tersebut masih berfokus pada klasifikasi satu tugas (single-task) dan belum menyentuh domain keluhan e-commerce secara langsung. Ihtada, Alfianita, dan Aziz menerapkan end-to-end IndoBERT, IndoBERT-CNN, dan IndoBERT-LSTM untuk klasifikasi multilabel ulasan e-commerce Indonesia, dengan IndoBERT-LSTM memberikan performa terbaik (F1-score 95,7%) dalam mengidentifikasi beberapa aspek ulasan secara simultan [23]. Namun, penelitian tersebut hanya mengklasifikasikan aspek ulasan (produk, layanan, pengiriman) tanpa mempertimbangkan tingkat prioritas penanganan setiap keluhan.

Beberapa penelitian melaporkan bahwa arsitektur hybrid yang mengombinasikan model transformer dengan BiLSTM mampu meningkatkan performa klasifikasi teks karena dapat menggabungkan representasi kontekstual dari transformer dengan kemampuan BiLSTM dalam memodelkan hubungan sekuensial antartoken [24], [25]. Meskipun IndoBERT telah memiliki mekanisme self-attention yang efektif dalam menangkap hubungan kontekstual antar token, beberapa penelitian menunjukkan bahwa penambahan lapisan BiLSTM masih dapat memberikan manfaat pada tugas klasifikasi teks. BiLSTM mampu memodelkan ketergantungan sekuensial secara eksplisit dari arah maju dan mundur sehingga dapat mempertahankan informasi urutan kata yang penting, terutama pada teks ulasan yang bersifat informal, mengandung kalimat panjang, serta memiliki variasi struktur bahasa. Dengan demikian, kombinasi IndoBERT sebagai pembentuk representasi kontekstual dan BiLSTM sebagai pemodel hubungan sekuensial diharapkan menghasilkan representasi fitur yang lebih komprehensif dibandingkan penggunaan model transformer saja.

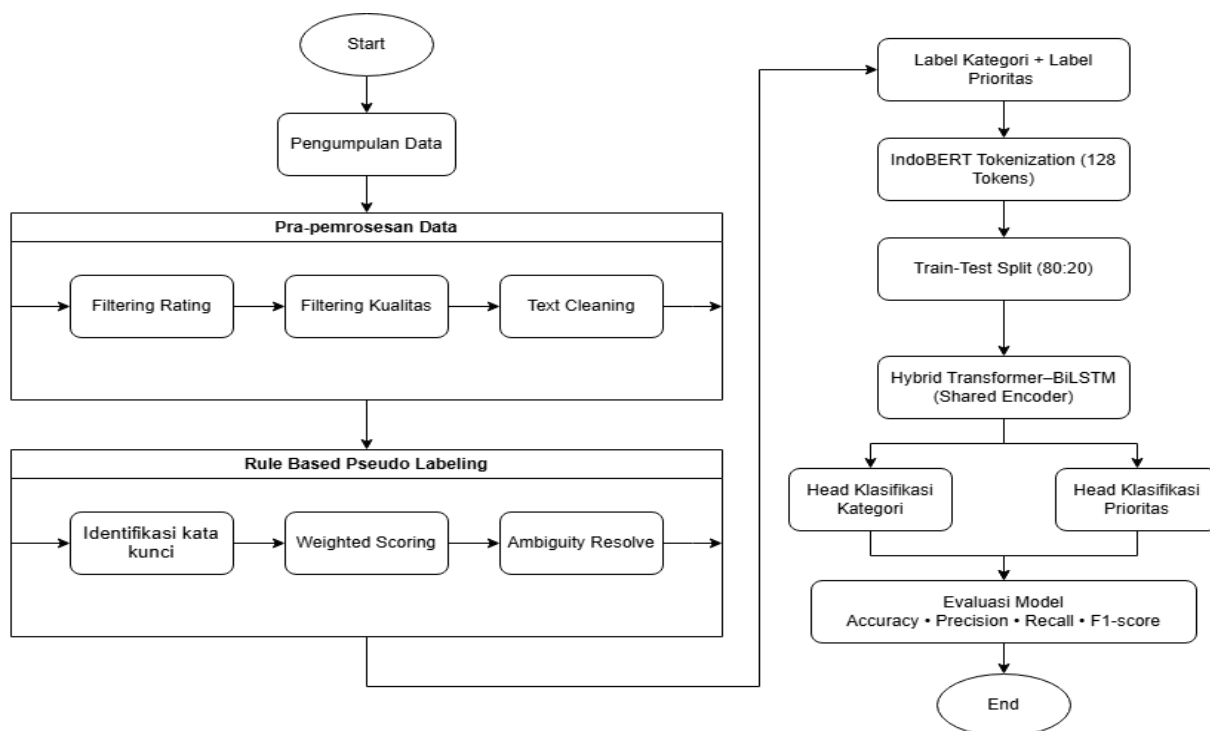
Pendekatan *multi-task learning* (MTL) semakin banyak diterapkan pada berbagai tugas klasifikasi karena mampu memanfaatkan keterkaitan antar tugas melalui mekanisme *shared representation* [26], [27]. Dengan berbagi representasi pada lapisan encoder, model dapat mempelajari fitur yang relevan untuk lebih dari satu tugas secara bersamaan sehingga meningkatkan efisiensi pembelajaran dan kemampuan generalisasi, terutama ketika tugas-tugas tersebut memiliki hubungan yang erat [28]. Klasifikasi kategori keluhan dan penentuan tingkat prioritas merupakan dua tugas yang saling berkaitan karena keduanya berasal dari teks ulasan yang sama. Informasi mengenai jenis keluhan dapat membantu model memahami tingkat urgensi penanganannya, begitu pula sebaliknya. Di sisi lain, Taneja, Vashishtha, dan Ratnoo menunjukkan bahwa DistilBERT mampu mempertahankan performa tinggi (F1-score 0,79–0,85) pada sentiment analysis dan product recommendation e-commerce meskipun data berdistribusi tidak seimbang, tanpa memerlukan teknik resampling [29]. Temuan ini relevan mengingat data keluhan pada penelitian kami juga mengalami ketidakseimbangan kelas, khususnya pada kategori prioritas menengah dan rendah, namun penelitian tersebut hanya berfokus pada dua tugas terpisah (sentimen dan rekomendasi) dan tidak dirancang dalam kerangka multi-task learning maupun domain keluhan. Oleh karena itu, pendekatan *multi-task learning* dipilih agar kedua tugas dapat memanfaatkan representasi fitur yang dipelajari secara bersama melalui *shared encoder*. Pendekatan ini diharapkan mampu meningkatkan efisiensi pembelajaran sekaligus menghasilkan representasi yang lebih informatif dibandingkan apabila kedua tugas dilatih secara terpisah. Berdasarkan empat penelitian terdahulu, masih terdapat beberapa kesenjangan penelitian yang perlu dikaji lebih lanjut. Penelitian oleh Aditama dan Wicaksono hanya berfokus pada klasifikasi keluhan dan non-keluhan sehingga belum mampu mengidentifikasi jenis keluhan maupun tingkat prioritas penanganannya [12]. Penelitian Zevana dan Riana menunjukkan efektivitas kombinasi IndoBERT

dengan CNN maupun BiLSTM, namun masih menerapkan pendekatan single-task pada analisis opini layanan pengiriman [22]. Selanjutnya, penelitian Ihtada, Alfianita, dan Aziz telah menerapkan IndoBERT-LSTM untuk klasifikasi multilabel aspek ulasan e-commerce, tetapi hanya mengidentifikasi aspek ulasan tanpa mempertimbangkan urgensi penanganan setiap keluhan [23]. Sementara itu, penelitian Taneja, Vashishtha, dan Ratnoo membuktikan bahwa model transformer mampu bekerja dengan baik pada data yang tidak seimbang, namun hanya digunakan untuk sentiment classification dan product recommendation secara terpisah, bukan untuk klasifikasi kategori dan prioritas dalam satu kerangka pembelajaran bersama [29]. Oleh karena itu, belum banyak penelitian yang mengintegrasikan arsitektur Hybrid IndoBERT–BiLSTM dengan pendekatan Multi-Task Learning untuk melakukan klasifikasi kategori dan prioritas keluhan pelanggan e-commerce berbahasa Indonesia secara simultan menggunakan satu shared encoder.

Berdasarkan uraian tersebut, penelitian ini mengusulkan sistem klasifikasi keluhan pelanggan Tokopedia menggunakan arsitektur Hybrid IndoBERT–BiLSTM dengan pendekatan *multi-task learning* untuk memprediksi kategori keluhan dan tingkat prioritas secara simultan. IndoBERT dimanfaatkan sebagai pembentuk representasi kontekstual bahasa Indonesia, sedangkan BiLSTM digunakan untuk memperkuat pemodelan hubungan sekuensial sebelum representasi fitur dibagikan kepada dua *classification head*. Kontribusi utama penelitian ini terletak pada penerapan pendekatan Hybrid IndoBERT–BiLSTM berbasis *multi-task learning* pada domain keluhan pelanggan e-commerce berbahasa Indonesia yang masih relatif terbatas dilaporkan dalam penelitian sebelumnya. Selain itu, penelitian ini menghasilkan dataset keluhan Tokopedia berbahasa Indonesia yang telah melalui proses pelabelan sehingga dapat dimanfaatkan sebagai referensi bagi penelitian lanjutan di bidang NLP dan manajemen layanan pelanggan digital.

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan *deep learning* dengan arsitektur hybrid Transformer-BiLSTM untuk klasifikasi multi-tugas keluhan pelanggan *e-commerce*. Metodologi penelitian terdiri dari delapan tahap utama: pengumpulan data, pra-pemrosesan data, *pseudo labeling*, tokenisasi, *data splitting*, pembangunan arsitektur model, *training*, dan evaluasi model. Diagram alur metodologi penelitian disajikan pada Gambar 1.



Gambar 1. Alur Penelitian

2.1 Pengumpulan Data

Dataset yang digunakan dalam penelitian ini merupakan kumpulan ulasan pelanggan aplikasi Tokopedia yang diperoleh melalui proses web scraping pada Google Play Store. Pemilihan Tokopedia sebagai objek penelitian didasarkan pada posisinya sebagai salah satu platform e-commerce terbesar di Indonesia dengan jutaan pengguna aktif yang secara konsisten memberikan umpan balik melalui ulasan aplikasi. Google Play Store dipilih sebagai sumber data karena menyediakan akses terbuka terhadap ulasan pengguna yang mencerminkan pengalaman nyata dalam menggunakan aplikasi, termasuk berbagai keluhan dan masalah yang dihadapi.

Proses pengumpulan data dilakukan menggunakan library Python *google-play-scraper* versi 1.2.4 untuk mengekstraksi ulasan pengguna dari Google Play Store melalui antarmuka publik yang disediakan oleh platform

tersebut. Penggunaan library ini memungkinkan proses pengambilan data dilakukan secara otomatis dengan tetap memanfaatkan informasi yang tersedia secara publik. Scraping dilakukan pada periode Desember 2024 hingga Januari 2025 untuk memastikan data yang diperoleh mencerminkan kondisi terkini dari aplikasi dan permasalahan yang sedang dihadapi pengguna. Proses scraping dirancang untuk mengambil seluruh komponen ulasan yang diperlukan, yaitu teks ulasan lengkap, rating bintang dalam skala 1 hingga 5, tanggal publikasi ulasan, nama pengguna yang memberikan ulasan, serta informasi metadata lainnya yang tersedia di platform. Data berhasil dikumpulkan sebanyak 100.000 ulasan dalam format data mentah yang disimpan dalam file CSV. Dataset mentah ini mencakup berbagai jenis ulasan mulai dari yang sangat positif (rating 5 bintang) hingga sangat negatif (rating 1 bintang), serta ulasan dengan panjang yang bervariasi mulai dari kalimat pendek hingga paragraf panjang. Data mentah ini juga mengandung berbagai karakteristik bahasa Indonesia informal yang umum digunakan di media sosial, seperti penggunaan kata tidak baku, singkatan, emoticon, serta campuran bahasa daerah.

Tabel 1 menunjukkan contoh data mentah yang berhasil diekstrak dari Google Play Store. Setiap baris merepresentasikan satu ulasan lengkap dengan rating dan timestamp yang nantinya akan digunakan untuk analisis temporal jika diperlukan. Terlihat bahwa data mentah masih sangat beragam, mulai dari ulasan negatif dengan rating rendah hingga ulasan positif dengan rating tinggi, serta mengandung berbagai elemen informal seperti emoticon, singkatan, dan bahasa gaul.

Tabel 1. Contoh data hasil scraping

Teks Ulasan	Rating	Tanggal
Aplikasi sering error saat mau checkout. Sudah berkali-kali coba masih gagal terus. Kecewa banget!	1	2024-12-15
Promo nya gak jelas, udah masukin kode voucher tapi gak ke apply. CS nya lama bgt responnya 😞	2	2024-12-18
Mantap aplikasinya, lancar jaya. Pengiriman cepat dan promo banyak. Recommended!	5	2024-12-25

2.2 Pra-Pemrosesan Data

Tahap pra-pemrosesan data merupakan langkah krusial dalam penelitian ini untuk mengubah data mentah menjadi data yang siap diproses oleh model deep learning. Pra-pemrosesan data dilakukan dalam tiga tahap utama yaitu filtering data berdasarkan rating, filtering ulasan berbobot berdasarkan kualitas konten, dan text cleaning untuk normalisasi teks. Tahap pertama preprocessing adalah melakukan filtering untuk menyaring ulasan yang benar-benar mengandung keluhan atau pengalaman negatif pengguna. Filtering berdasarkan rating bintang dilakukan dengan hanya mempertahankan ulasan dengan rating 1 hingga 3 bintang. Keputusan threshold ini didasarkan pada asumsi bahwa pengguna yang memberikan rating rendah cenderung menyampaikan kritik, keluhan, atau masalah yang mereka hadapi saat menggunakan aplikasi. Ulasan dengan rating 4 dan 5 bintang dieliminasi karena umumnya berisi testimoni positif, pujian, atau pengalaman memuaskan yang tidak termasuk dalam kategori keluhan. Proses filtering rating ini mengurangi dataset dari 100.000 ulasan mentah menjadi 44.128 ulasan.

Tahap filtering kedua dilakukan untuk menghapus ulasan yang tidak memiliki kualitas informasi yang memadai untuk proses klasifikasi. Kriteria yang digunakan meliputi panjang teks minimum, kebermaknaan konten, deteksi noise, deteksi spam, serta penghapusan duplikasi. Ulasan dengan panjang kurang dari lima kata tidak secara otomatis dihapus hanya karena jumlah katanya yang sedikit, tetapi setelah dilakukan pemeriksaan ditemukan bahwa sebagian besar ulasan tersebut hanya berupa ungkapan umum seperti "*jelek*", "*buruk*", "*kecewa*", atau "*bagus*" tanpa memberikan informasi mengenai penyebab maupun jenis keluhan. Oleh karena itu, kriteria panjang minimum lima kata digunakan sebagai salah satu indikator kualitas informasi untuk membantu mempertahankan ulasan yang memiliki konteks yang lebih memadai bagi proses klasifikasi. Meskipun demikian, penetapan batas ini berpotensi menghilangkan sebagian kecil ulasan singkat yang masih mengandung informasi penting dan menjadi salah satu keterbatasan penelitian yang akan dipertimbangkan pada penelitian selanjutnya. Ulasan yang hanya berisi simbol, emoticon, atau karakter berulang tanpa makna linguistik dihapus karena tidak memberikan informasi substantif. Ulasan yang teridentifikasi sebagai spam atau tidak relevan dengan fungsi aplikasi, seperti promosi aplikasi lain, juga dieliminasi. Untuk menghindari bias akibat pengulangan data, dilakukan penghapusan duplikasi menggunakan pendekatan exact match pada teks ulasan. Setelah tahap ini, jumlah data berkurang dari 44.128 menjadi 42.929 ulasan yang dianggap berkualitas dan relevan untuk digunakan dalam proses analisis selanjutnya.

Setelah melalui tahap filtering, data ulasan yang tersisa masih mengandung berbagai elemen noise yang berpotensi mengganggu proses pembelajaran model. Oleh karena itu, dilakukan tahap *text cleaning* dan normalisasi untuk menghasilkan teks yang lebih terstruktur dan konsisten. Proses ini diawali dengan *lowercasing*, yaitu konversi seluruh teks menjadi huruf kecil untuk menghilangkan variasi *case sensitivity*. Selanjutnya, dilakukan normalisasi kata tidak baku menggunakan kamus khusus yang memetakan kata informal ke bentuk baku, mengingat karakteristik ulasan pengguna yang cenderung menggunakan bahasa tidak formal.

Selain itu, elemen non-informatif seperti URL, tautan, *mention* (@username), dan *hashtag* dihapus karena tidak berkontribusi terhadap makna keluhan. Tanda baca berlebihan dinormalisasi untuk mengurangi variasi yang tidak diperlukan, dan spasi ganda disederhanakan menjadi satu spasi. Emoticon serta emoji juga dihapus karena informasi

sentimen telah direpresentasikan melalui rating numerik. Sebagai ilustrasi, teks seperti “Aplikasi error bgttttt!!! Gak bisa checkout 🤔” dinormalisasi menjadi “aplikasi error banget tidak bisa checkout”. Seluruh proses pembersihan hanya menghapus elemen yang tidak memberikan kontribusi terhadap informasi semantik, sedangkan isi utama keluhan tetap dipertahankan agar representasi yang dipelajari model tetap mencerminkan konteks asli ulasan pelanggan.

2.3 Pseudo Labeling

Setelah seluruh proses preprocessing dilakukan, diperoleh dataset sebanyak 42.929 ulasan yang bersih, konsisten, dan siap digunakan untuk tahap pseudo labeling. Dataset ini memiliki karakteristik bahasa yang lebih terstandarisasi sehingga dapat meningkatkan efektivitas proses pembelajaran model deep learning. Pendekatan pseudo labeling dalam penelitian ini terdiri atas tiga komponen utama, yaitu identifikasi kata kunci, formulasi skoring deterministik, dan mekanisme resolusi ambiguitas. Pada tahap identifikasi kata kunci, dilakukan analisis terhadap 500 data yang dipilih secara acak dari 42.929 sampel ulasan untuk menginventarisasi kata atau frasa yang paling sering muncul untuk dua tujuan pelabelan, yaitu komponen kategori, dan prioritas keluhan (prioritas). Setiap kata kunci dievaluasi berdasarkan tiga dimensi, yaitu dampak (*impact*), kepastian negatif (*polarity certainty*), dan spesifisitas (*specificity*). Ketiga dimensi tersebut digunakan untuk menentukan bobot secara deterministik dengan nilai 1 atau 2. Berdasarkan proses ini, dari 500 data yang digunakan diperoleh 327 kata kunci yang unik, yang terdiri atas 241 kata kunci untuk pelabelan pada komponen kategori dan 86 kata kunci untuk pelabelan pada prioritas. Pelabelan dilakukan menggunakan aturan deterministik berbasis kata kunci sehingga setiap ulasan dengan karakteristik yang sama akan memperoleh label yang konsisten. Penelitian ini tidak menerapkan validasi human-in-the-loop maupun inter-annotator agreement karena kedua metode tersebut tidak termasuk dalam ruang lingkup penelitian.

2.3.1 Pelabelan Ulasan Pada Komponen Kategori

Tahap ini bertujuan mentransformasikan ulasan berbentuk teks bebas menjadi label terstruktur yang digunakan sebagai target pada pelatihan model klasifikasi. Penentuan kategori dilakukan melalui skoring pada 241 kata kunci berbasis frekuensi, dengan menghitung skor tiap kategori dari frekuensi kemunculan kata kunci relevan yang dikalikan bobotnya. Secara matematis, skor kategori dihitung sebagai akumulasi kontribusi setiap kata kunci yang terdeteksi pada kelima kategori tersebut, sebagaimana dirumuskan pada Persamaan (1):

$$\text{score}_k = \sum_{i=1}^{n_k} \text{freq}(k w_{k,i}) \times w_{k,i} \quad (1)$$

dengan $\text{freq}(k w_{k,i})$ menyatakan frekuensi kemunculan kata kunci ke- i pada kategori ke- k , $w_{k,i}$ menyatakan bobot kata kunci, dan n_k menyatakan jumlah kata kunci pada kategori ke- k .

Label kategori kemudian ditetapkan menggunakan fungsi argmax, yaitu memilih kategori dengan skor tertinggi, sebagaimana pada Persamaan (2):

$$\hat{y}_{\text{kategori}} = \arg \max_k (\text{score}_k) \quad (2)$$

Apabila dua atau lebih kategori memiliki skor identik, digunakan mekanisme tie-breaking berbasis prioritas domain dengan urutan: Transaksi, Pengiriman, Teknis Aplikasi, Promo, dan Pelayanan CS. Urutan ini ditetapkan berdasarkan tingkat dampak permasalahan pada konteks e-commerce, di mana isu yang berkaitan dengan kerugian finansial dan gangguan operasional dipandang memiliki urgensi lebih tinggi. Dengan mekanisme tersebut, proses pelabelan kategori dapat dilakukan secara sistematis, konsisten, dan dapat direproduksi pada seluruh dataset penelitian. Pelabelan menghasilkan lima komponen kategori: Transaksi, Pengiriman, Teknis Aplikasi, Promo, dan Pelayanan CS, yang untuk selanjutnya masing-masing disingkat dengan kategori A, B, C, D dan E. Proses ini menghasilkan distribusi pada 241 kata kunci ke dalam masing-masing komponen kategori tersebut sejumlah 42, 53, 54, 35 dan 57. Tabel 2 menguraikan hasil distribusi dalam setiap komponen kategori berserta contoh kata kuncinya.

Tabel 2. Contoh kata kunci setiap kategori keluhan

Kategori	Jumlah	Contoh Kata kunci
A	42	checkout, bayar, pembayaran, refund, saldo, transfer, gagal bayar, metode pembayaran, gopay, ovo
B	53	kirim, pengiriman, terlambat, rusak, salah alamat, tracking, kurir, paket, belum sampai
C	54	error, bug, crash, loading, lemot, hang, gagal, tidak berfungsi, rusak, force close
D	35	voucher, promo, diskon, cashback, kode promo, kupon, tidak dapat digunakan, expired
E	57	cs, customer service, komplain, laporan, tidak ada respon, lama balas, admin, chat

2.3.2 Labeling Prioritas Keluhan

Komponen prioritas dibagi menjadi tiga kelas, yaitu Tinggi, Sedang, dan Rendah. Tahap ini bertujuan mentransformasikan tingkat keparahan keluhan menjadi label terstruktur yang digunakan sebagai target kedua dalam skema *multi-task learning*. Penentuan prioritas untuk setiap ulasan dilakukan menggunakan pendekatan skoring multi-

kriteria yang menggabungkan intensitas kata kunci, tingkat keparahan berdasarkan rating, dan tingkat kedetailan deskripsi teks. Skor prioritas ulasan dihitung sebagai penjumlahan dari ketiga komponen tersebut, sebagaimana dirumuskan pada Persamaan (3):

$$\text{priority}_{\text{score}} = \text{score}_{\text{keyword}} + \text{score}_{\text{rating}} + \text{score}_{\text{length}} \quad (3)$$

dengan $\text{score}_{\text{keyword}}$ menyatakan akumulasi bobot kata kunci prioritas yang terdeteksi, $\text{score}_{\text{rating}}$ menyatakan kontribusi tingkat keparahan berdasarkan rating bintang, dan $\text{score}_{\text{length}}$ menyatakan kontribusi informasi berdasarkan panjang teks ulasan. Komponen skor berdasarkan rating ditetapkan dengan memberikan nilai lebih tinggi pada rating yang lebih rendah, karena diasumsikan merepresentasikan tingkat ketidakpuasan yang lebih besar. Rating 1 diberikan skor 3, rating 2 diberikan skor 2, dan rating 3 diberikan skor 1. Sementara itu, komponen panjang teks digunakan sebagai penguat konteks, di mana ulasan dengan panjang lebih dari 200 kata diberi skor 1, ulasan dengan panjang 50–200 kata diberi skor 0,5, dan ulasan dengan panjang kurang dari 50 kata diberi skor 0.

Skor akhir kemudian dipetakan menjadi label prioritas menggunakan ambang keputusan sebagaimana dirumuskan pada Persamaan (4):

$$\hat{y}_{\text{prioritas}} = \begin{cases} \text{Tinggi,} & \text{jika } \text{priority}_{\text{score}} \geq 5 \\ \text{Sedang,} & \text{jika } 3 \leq \text{priority}_{\text{score}} < 5 \\ \text{Rendah,} & \text{jika } \text{priority}_{\text{score}} < 3 \end{cases} \quad (4)$$

Selain mekanisme berbasis skoring, diterapkan pula mekanisme *override* untuk menangani kasus-kasus kritis. Apabila terdeteksi kata kunci yang berkaitan dengan fraud, kehilangan finansial, kompromi akun, atau keamanan data, maka label prioritas secara otomatis ditetapkan sebagai prioritas Tinggi meskipun skor numerik berada pada rentang kelas lain. Mekanisme ini diterapkan untuk menjaga sensitivitas terhadap kasus berisiko tinggi yang memerlukan penanganan segera. Dengan pendekatan tersebut, proses pelabelan prioritas dapat dilakukan secara sistematis, konsisten, deterministik, dan responsif terhadap kondisi khusus pada seluruh dataset penelitian. Proses ini menghasilkan distribusi pada 86 kata kunci ke dalam masing-masing komponen prioritas tersebut sejumlah 37, 33, dan 16. Tabel 3 menguraikan hasil distribusi dalam setiap komponen prioritas berserta contoh kata kuncinya.

Tabel 3. Contoh kata kunci setiap tingkat prioritas

Prioritas	Jumlah	Contoh Kata kunci
Tinggi	37	penipuan, scam, uang hilang, tidak bisa akses, data bocor, hacker, sangat kecewa, terburuk, jangan install, hapus aplikasi, bohong, menipu, fatal error
Sedang	33	kecewa, harap diperbaiki, tolong perbaiki, perlu ditingkatkan, kurang memuaskan, sering error, menunggu lama, bermasalah, tidak nyaman
Rendah	16	saran, sebaiknya, alangkah baiknya, bisa ditambahkan, mungkin perlu, lumayan, cukup baik tapi, perlu improve, minor bug

2.3.3 Distribusi Hasil Pelabelan Dataset

Distribusi hasil pelabelan digunakan untuk menggambarkan karakteristik data penelitian sekaligus memberikan gambaran mengenai keseimbangan kelas (*class distribution*) pada dataset yang akan digunakan dalam proses pelatihan model. Sebelum dilakukan analisis distribusi data, Tabel 4 menunjukkan beberapa contoh hasil pseudo labeling yang dihasilkan secara otomatis menggunakan pendekatan berbasis aturan (*rule-based labeling*). Setiap ulasan memperoleh dua jenis label, yaitu kategori dan prioritas, berdasarkan hasil perhitungan skor kata kunci, rating, serta konteks isi ulasan.

Tabel 4. Contoh data hasil pseudo labeling

Kategori	Teks Ulasan	Prioritas
Transaksi	pembayaran sudah berhasil tapi status pesanan tidak berubah dan dana belum kembali, sangat mengecewakan	Tinggi
Transaksi	checkout kadang gagal saat jam promo, mohon perbaiki karena cukup mengganggu	Rendah
Pengiriman	paket belum sampai lebih dari seminggu dan status resi tidak bergerak, mohon ditindak	Tinggi
Pengiriman	barang diterima terlambat satu hari dari estimasi, selebihnya masih cukup baik	Rendah
Teknis Aplikasi	aplikasi sering crash saat buka menu pembayaran dan tidak bisa lanjut checkout	Tinggi
Teknis Aplikasi	loading halaman produk agak lama di beberapa waktu, semoga performa ditingkatkan	Rendah
Promo	voucher cashback tidak bisa dipakai padahal syarat sudah terpenuhi, saya kecewa	Sedang
Pelayanan CS	respon cs agak lama namun akhirnya masalah terselesaikan, ke depan semoga lebih cepat	Sedang

Distribusi kategori menunjukkan bahwa dari 42.929 ulasan, keluhan terkait Transaksi memiliki jumlah tertinggi sebanyak 14.189 ulasan atau sekitar 33,05% dari total dataset. Selanjutnya, kategori Pelayanan CS memiliki 10.578 ulasan (24,64%), Pengiriman sebanyak 9.635 ulasan (22,44%), Teknis Aplikasi sebanyak 7.801 ulasan (18,17%), dan Promo sebanyak 726 ulasan (1,69%). Distribusi tersebut menunjukkan bahwa mayoritas permasalahan pengguna berkaitan dengan proses transaksi dan layanan pelanggan. Sementara itu, kategori Promo memiliki jumlah data yang relatif sedikit dibandingkan kategori lainnya, sehingga menghasilkan distribusi kelas yang tidak seimbang (*imbalanced dataset*). Kondisi ini umum ditemukan pada data ulasan pengguna di platform e-commerce karena tidak semua pengguna mengalami atau membahas kendala promo dalam ulasannya. Tabel 5 menunjukkan distribusi hasil pelabelan kategori pada dataset penelitian.

Tabel 5. Distribusi label kategori

Kategori	Jumlah Ulasan	Persentase
Transaksi	14.189	33,05%
Pengiriman	9.635	22,44%
Teknis Aplikasi	7.801	18,17%
Promo	726	1,69%
Pelayanan CS	10.578	24,64%

Selain label kategori, proses pelabelan juga menghasilkan distribusi label prioritas yang terdiri atas prioritas Tinggi, Sedang, dan Rendah. Hasil distribusi menunjukkan bahwa dari 42.929 ulasan, prioritas Tinggi mendominasi dataset dengan jumlah 34.219 ulasan atau sekitar 79,71% dari total data. Prioritas Sedang berjumlah 6.166 ulasan (14,36%), sedangkan prioritas Rendah berjumlah 2.544 ulasan (5,93%). Dominasi prioritas Tinggi menunjukkan bahwa sebagian besar ulasan pengguna mengandung indikasi keluhan serius, seperti gangguan transaksi, kendala akses aplikasi, keterlambatan pengiriman, maupun masalah pelayanan yang memerlukan penanganan cepat. Distribusi ini juga mengindikasikan adanya ketidakseimbangan kelas pada label prioritas, khususnya antara prioritas Tinggi dan Rendah. Tabel 6 menunjukkan distribusi label prioritas pada dataset penelitian.

Tabel 6. Distribusi label prioritas

Prioritas	Jumlah Ulasan	Persentase
Tinggi	34.219	79,71%
Sedang	6.166	14,36%
Rendah	2.544	5,93%

Distribusi prioritas pada masing-masing kategori juga menunjukkan pola yang relatif konsisten, di mana sebagian besar ulasan pada setiap kategori termasuk ke dalam prioritas Tinggi. Kondisi tersebut menunjukkan bahwa pengguna cenderung memberikan ulasan ketika mengalami permasalahan yang berdampak signifikan terhadap pengalaman penggunaan aplikasi. Tabel 7 menunjukkan distribusi gabungan antara kategori dan tingkat prioritas keluhan.

Tabel 7. Distribusi prioritas pada setiap kategori

Kategori	Tinggi	Sedang	Rendah
Transaksi	11.620	1.755	814
Pengiriman	7.592	1.668	375
Teknis Aplikasi	5.949	1.376	476
Promo	596	96	34
Pelayanan CS	8.462	1.271	845

Berdasarkan distribusi tersebut, dapat disimpulkan bahwa dataset penelitian memiliki karakteristik *imbalanced multi-class dataset*, baik pada komponen kategori maupun prioritas. Oleh karena itu, pada tahap pelatihan model diperlukan penggunaan metrik evaluasi yang tidak hanya berfokus pada accuracy, tetapi juga mempertimbangkan precision, recall, dan F1-score agar performa model pada kelas minoritas tetap dapat dievaluasi secara optimal.

2.4 Implementasi Model

Pada tahap ini, data ulasan yang telah memiliki label kategori dan prioritas dipersiapkan untuk proses pelatihan model berbasis transformer menggunakan IndoBERT. Tahapan implementasi meliputi proses tokenisasi teks, pembagian dataset menjadi data pelatihan dan validasi, serta penanganan distribusi kelas sebelum dilakukan proses pelatihan model.

2.4.1 Tokenisasi

Tahap tokenisasi bertujuan mengubah data ulasan berbentuk teks menjadi representasi numerik yang dapat diproses oleh model *deep learning*. Pada penelitian ini digunakan tokenizer dari model IndoBERT, yaitu *indobenchmark/indobert-base-pl*, yang telah dilatih menggunakan korpus besar berbahasa Indonesia sehingga

mampu memahami konteks linguistik bahasa Indonesia dengan lebih baik dibandingkan pendekatan tokenisasi konvensional.

Tokenizer IndoBERT menggunakan metode *subword tokenization* berbasis WordPiece. Pendekatan ini memungkinkan kata yang tidak terdapat dalam kosakata (*out-of-vocabulary*) tetap dapat direpresentasikan melalui pemecahan menjadi subword yang lebih kecil. Dengan demikian, model tetap mampu memahami variasi kata tidak baku, singkatan, maupun istilah khas yang umum ditemukan pada ulasan pengguna aplikasi e-commerce. Tabel 8 menunjukkan spesifikasi tokenizer yang digunakan pada penelitian ini.

Tabel 8. Spesifikasi tokenizer IndoBERT

Komponen	Spesifikasi
Model	Indobenchmark/indober-base-p1
Tipe Tokenizer	Subword tokenizer (WordPiece)
Vocabulary Size	±60.000 token
Maximum Sequence Length	128 token
Special Tokens	[CLS], [SEP], [UNK], [PAD]

Pada proses tokenisasi, setiap ulasan dikonversi menjadi urutan token dengan panjang maksimum 128 token. Token khusus [CLS] ditempatkan pada awal kalimat sebagai representasi keseluruhan sekuens, sedangkan token [SEP] digunakan sebagai penanda akhir kalimat. Apabila jumlah token melebihi batas maksimum, maka dilakukan proses *truncation* untuk memotong token yang berada di bagian akhir. Sebaliknya, apabila panjang token kurang dari 128 token, maka dilakukan *padding* menggunakan token [PAD] hingga seluruh data memiliki panjang sekuens yang seragam.

Penelitian ini menggunakan panjang urutan maksimum (*maximum sequence length*) sebesar 128 token dengan mekanisme *truncation* untuk ulasan yang melebihi batas tersebut. Pemilihan nilai ini didasarkan pada karakteristik dataset, di mana sebagian besar ulasan pelanggan memiliki panjang di bawah 128 token sehingga dapat direpresentasikan secara utuh. Ulasan pada Google Play Store umumnya menyampaikan inti permasalahan pada bagian awal hingga tengah teks. Dengan demikian, penggunaan *max length* sebesar 128 diharapkan mampu mempertahankan informasi yang paling relevan sekaligus menjaga efisiensi komputasi selama proses pelatihan model.

Proses tokenisasi dilakukan terhadap seluruh 42.929 ulasan yang telah memiliki label kategori dan prioritas. Hasil tokenisasi menghasilkan representasi data berbentuk matriks token dengan dimensi (42.929, 128) yang selanjutnya digunakan sebagai masukan utama pada model klasifikasi berbasis transformer. Selain menghasilkan *input_ids*, proses tokenisasi juga menghasilkan *attention mask* yang digunakan untuk membedakan token aktual dan token hasil *padding*. Mekanisme ini membantu model dalam memfokuskan perhatian (*attention*) hanya pada token yang memiliki informasi relevan selama proses pelatihan.

2.4.2 Pembagian Data Pelatihan dan Validasi

Distribusi data pelatihan hasil stratifikasi menunjukkan bahwa kategori Transaksi tetap mendominasi dengan 11.351 data, terdiri atas 9.296 prioritas Tinggi, 1.404 prioritas Sedang, dan 651 prioritas Rendah. Kategori Pengiriman memiliki 7.708 data, Teknis Aplikasi sebanyak 6.241 data, Pelayanan CS sebanyak 8.462 data, dan Promo sebanyak 581 data. Secara keseluruhan, data pelatihan terdiri atas 27.375 prioritas Tinggi, 4.933 prioritas Sedang, dan 2.035 prioritas Rendah dari total 34.343 ulasan.

Sementara itu, distribusi data validasi juga mempertahankan proporsi kelas yang serupa dengan dataset awal. Kategori Transaksi pada data validasi berjumlah 2.838 data yang terdiri atas 2.324 prioritas Tinggi, 351 prioritas Sedang, dan 163 prioritas Rendah. Kategori Pengiriman memiliki 1.927 data, Teknis Aplikasi sebanyak 1.560 data, Pelayanan CS sebanyak 2.116 data, dan Promo sebanyak 145 data. Secara keseluruhan, data validasi terdiri atas 6.844 prioritas Tinggi, 1.233 prioritas Sedang, dan 509 prioritas Rendah dari total 8.586 ulasan. Berdasarkan distribusi tersebut, dapat dilihat bahwa proporsi masing-masing kelas pada data pelatihan dan data validasi tetap konsisten terhadap distribusi dataset awal. Hal ini menunjukkan bahwa proses *stratified splitting* berhasil mempertahankan keseimbangan distribusi kelas sehingga hasil evaluasi model dapat merepresentasikan performa model secara lebih objektif.

2.4.3 Arsitektur Model Hybrid Transformer-BiLSTM

Penelitian ini menggunakan arsitektur *Hybrid Transformer-BiLSTM* dengan pendekatan *multi-task learning* untuk melakukan klasifikasi kategori dan prioritas keluhan secara simultan. Model menggunakan indobenchmark/indobert-base-p1 sebagai *transformer encoder* untuk menghasilkan representasi kontekstual teks, kemudian dilanjutkan dengan lapisan BiLSTM berukuran 128 unit dua arah untuk menangkap dependensi sekuensial antar token. Representasi fitur hasil BiLSTM diproses melalui *shared dense layer* berukuran 256 unit dengan aktivasi ReLU sebelum diteruskan ke dua *classification head*. *Category head* digunakan untuk memprediksi lima kategori keluhan menggunakan fungsi aktivasi Softmax, sedangkan *priority head* digunakan untuk memprediksi tiga tingkat prioritas. Model memiliki total 125.492.488 parameter dan seluruh parameter bersifat *trainable*.

2.4.4 Konfigurasi Pelatihan Model

Pelatihan model dilakukan menggunakan optimizer Adam dengan *learning rate* sebesar 2×10^{-5} , *batch size* 16, dan maksimum 10 epoch. Untuk mencegah *overfitting*, diterapkan teknik *dropout* sebesar 0,3 pada lapisan BiLSTM, *shared dense layer*, dan *classification head*. Fungsi *loss* yang digunakan adalah *categorical cross-entropy* untuk kedua tugas klasifikasi dengan pembobotan yang seimbang antara tugas kategori dan prioritas.

$$L_{total} = 1.0 \times L_{category} + 1.0 \times L_{priority} \quad (5)$$

Selain itu, digunakan mekanisme *early stopping* dengan *patience* 3 epoch dan *ReduceLROnPlateau* dengan faktor penurunan 0,5 untuk menjaga stabilitas proses pelatihan model.

2.5 Evaluasi Model

Evaluasi model dilakukan menggunakan metrik Accuracy, Precision, Recall, dan F1-score. Accuracy digunakan untuk mengukur proporsi prediksi yang benar terhadap seluruh data uji. Precision menunjukkan tingkat ketepatan prediksi pada setiap kelas, Recall mengukur kemampuan model dalam mengenali seluruh data pada suatu kelas, sedangkan F1-score merupakan rata-rata harmonis antara Precision dan Recall. Evaluasi dilakukan pada kedua tugas, yaitu klasifikasi kategori dan klasifikasi prioritas. Secara umum, model Hybrid Transformer-BiLSTM menunjukkan performa yang sangat baik pada klasifikasi kategori keluhan. Kombinasi IndoBERT dan BiLSTM mampu memahami konteks semantik dan pola sekuensial pada ulasan pengguna berbahasa Indonesia secara efektif. Sementara itu, performa klasifikasi prioritas masih dipengaruhi oleh ketidakseimbangan distribusi kelas, di mana prioritas Tinggi mendominasi dataset. Meskipun demikian, model tetap mampu mendeteksi keluhan prioritas Tinggi dengan baik, sehingga berpotensi membantu proses penanganan keluhan yang memerlukan respons cepat pada platform e-commerce.

3. HASIL DAN PEMBAHASAN

3.1 Hasil

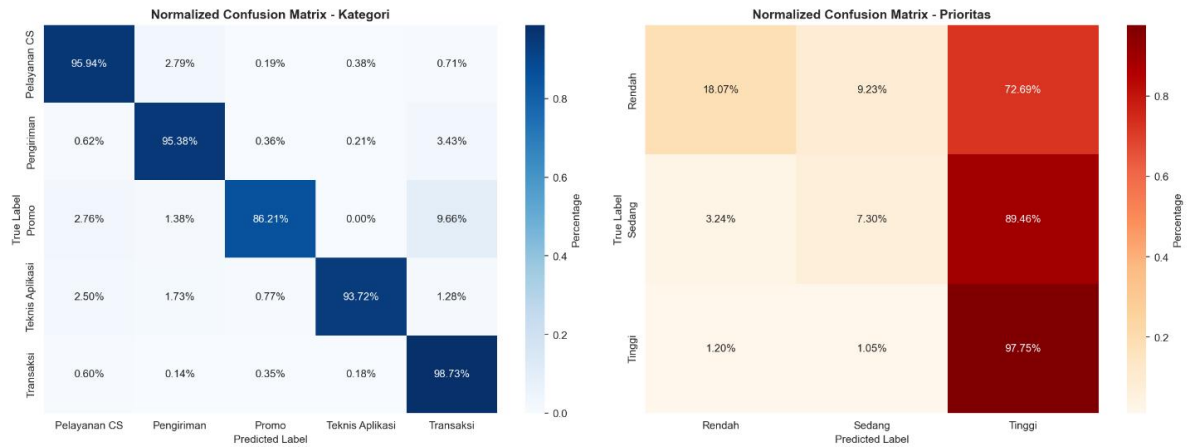
Penelitian ini menghasilkan model klasifikasi keluhan pelanggan Tokopedia berbasis *Hybrid Transformer-BiLSTM* dengan pendekatan *multi-task learning*. Model dirancang untuk melakukan dua tugas klasifikasi secara simultan, yaitu klasifikasi kategori keluhan dan klasifikasi prioritas keluhan. Seluruh proses implementasi dilakukan menggunakan dataset ulasan aplikasi Tokopedia yang diperoleh dari Google Play Store dan telah melalui tahapan *preprocessing* serta *pseudo labeling*. Tahap *preprocessing* berhasil mengubah 100.000 data mentah menjadi 42.929 ulasan yang dinilai relevan dan berkualitas untuk proses pelatihan model.

Proses tersebut meliputi filtering rating, filtering kualitas ulasan, serta normalisasi teks. Hasil akhir preprocessing menghasilkan data yang lebih konsisten dan representatif terhadap konteks keluhan pengguna. Selanjutnya, seluruh ulasan ditokenisasi menggunakan tokenizer IndoBERT (*indobenchmark/indobert-base-pl*) dengan panjang maksimum 128 token. Proses tokenisasi menghasilkan representasi numerik berupa *input_ids* dan *attention mask* yang digunakan sebagai masukan model. Dataset kemudian dibagi menjadi data pelatihan dan data validasi menggunakan metode *stratified train-test split* dengan rasio 80:20.

Teknik stratifikasi digunakan untuk mempertahankan proporsi distribusi kelas pada kedua himpunan data sehingga evaluasi model dapat dilakukan secara lebih objektif. Model yang digunakan mengombinasikan IndoBERT sebagai *transformer encoder* dan BiLSTM sebagai lapisan sekuensial. IndoBERT digunakan untuk menangkap representasi kontekstual teks berbahasa Indonesia, sedangkan BiLSTM digunakan untuk mempelajari hubungan antar token secara dua arah. Representasi fitur selanjutnya diteruskan ke *shared dense layer* sebelum diproses oleh dua *classification head*, yaitu *category head* dan *priority head*. Pelatihan model dilakukan menggunakan optimizer Adam dengan *learning rate* sebesar 2×10^{-5} , *batch size* 16, dan maksimum 10 epoch. Untuk mencegah *overfitting*, diterapkan teknik *dropout* sebesar 0,3 pada beberapa lapisan model serta mekanisme *early stopping*.

Gambar 2 menunjukkan *normalized confusion matrix* pada tugas klasifikasi kategori dan klasifikasi prioritas. Pada klasifikasi kategori, sebagian besar prediksi berada pada diagonal utama yang menunjukkan bahwa model mampu mengklasifikasikan mayoritas ulasan ke kategori yang sesuai. Tingkat klasifikasi benar tertinggi dicapai pada kategori Transaksi sebesar 98,73%, sedangkan kategori Promo memiliki tingkat klasifikasi benar sebesar 86,21%, yang merupakan nilai terendah dibandingkan kategori lainnya. Sementara itu, pada klasifikasi prioritas terlihat bahwa model mampu mengenali kelas Tinggi dengan sangat baik, ditunjukkan oleh tingkat klasifikasi benar sebesar 97,75%. Sebaliknya, kelas Sedang dan Rendah menunjukkan tingkat klasifikasi benar yang lebih rendah serta lebih sering diprediksi sebagai kelas Tinggi, yang mengindikasikan adanya pengaruh ketidakseimbangan distribusi data terhadap kemampuan model dalam membedakan tingkat urgensi keluhan.

Hasil pengujian menunjukkan bahwa model memiliki performa yang sangat baik pada tugas klasifikasi kategori keluhan. Model memperoleh akurasi sebesar 96,17% dengan *weighted F1-score* sebesar 0,9617 dan *macro F1-score* sebesar 0,9353. Nilai tersebut menunjukkan bahwa model mampu mengenali sebagian besar kategori keluhan secara konsisten. Tabel 9 menunjukkan hasil evaluasi klasifikasi kategori pada data uji.



Gambar 2. Normalized Confusion Matrix Klasifikasi Kategori dan Prioritas

Tabel 9. Hasil Evaluasi Klasifikasi Kategori

Kategori	Precision	Recall	F1-Score
Transaksi	0,9606	0,9873	0,9738
Pengiriman	0,9523	0,9538	0,9531
Teknis Aplikasi	0,9885	0,9372	0,9622
Promo	0,7911	0,8621	0,8251
Pelayanan CS	0,9657	0,9594	0,9625

Berdasarkan hasil evaluasi, kategori Promo memperoleh nilai precision sebesar 0,7911, recall sebesar 0,8621, dan F1-score sebesar 0,8251, yang merupakan nilai terendah dibandingkan kategori lainnya. Salah satu faktor yang diduga memengaruhi kondisi tersebut adalah ketidakseimbangan distribusi data, di mana kategori Promo hanya terdiri atas 726 ulasan (1,69%) dari total 42.929 data. Jumlah sampel yang relatif sedikit menyebabkan variasi pola bahasa yang dipelajari model menjadi lebih terbatas dibandingkan kategori mayoritas, sehingga representasi fitur yang dibentuk model untuk kategori Promo belum sekuat kelas lainnya.

Nilai recall yang mencapai 0,8621 menunjukkan bahwa sebagian besar data Promo masih mampu dikenali dengan baik oleh model. Selain itu, nilai F1-score sebesar 0,8251 mengindikasikan bahwa performa klasifikasi pada kategori minoritas masih tergolong cukup baik dan belum menunjukkan indikasi terjadinya *minority class collapse*, yaitu kondisi ketika model gagal mengenali kelas minoritas akibat dominasi kelas mayoritas. Dengan demikian, meskipun terdapat pengaruh ketidakseimbangan data terhadap performa kategori Promo, dampaknya masih berada pada tingkat yang dapat ditoleransi dalam proses klasifikasi.

Temuan tersebut juga konsisten dengan Gambar 2, yang menunjukkan bahwa sebagian besar data pada kategori Promo tetap berhasil diklasifikasikan dengan benar (86,21%), meskipun masih terdapat sebagian kecil ulasan yang diprediksi sebagai kategori lain. Hal ini menunjukkan bahwa ketidakseimbangan distribusi data memengaruhi performa model, tetapi belum menyebabkan penurunan kemampuan klasifikasi secara signifikan.

Pada tugas klasifikasi prioritas, model memperoleh akurasi sebesar 80,04% dengan *weighted F1-score* sebesar 0,7437 dan *macro F1-score* sebesar 0,4236. Hasil ini menunjukkan bahwa tugas klasifikasi prioritas memiliki tingkat kesulitan yang lebih tinggi dibandingkan klasifikasi kategori. Tabel 10 menunjukkan hasil evaluasi klasifikasi prioritas pada data uji.

Tabel 10. Hasil Evaluasi Klasifikasi Prioritas

Prioritas	Precision	Recall	F1-Score
Tinggi	0,8196	0,9775	0,8916
Sedang	0,4306	0,0730	0,1248
Rendah	0,4299	0,1807	0,2545

Berdasarkan hasil evaluasi pada Tabel 10, model menunjukkan performa yang sangat baik pada kelas Prioritas Tinggi dengan nilai precision sebesar 0,8196, recall sebesar 0,9775, dan F1-score sebesar 0,8916. Tingginya nilai recall menunjukkan bahwa hampir seluruh keluhan yang memiliki tingkat urgensi tinggi berhasil dikenali oleh model. Kemampuan ini menjadi penting dalam implementasi sistem penanganan keluhan otomatis karena dapat meminimalkan risiko terlewatnya keluhan yang memerlukan respons segera.

Performa pada kelas prioritas Sedang dan Rendah masih relatif rendah, terutama pada nilai recall yang hanya sebesar 0,0730 dan 0,1807. Kondisi ini menunjukkan bahwa model masih mengalami kesulitan dalam membedakan tingkat urgensi pada kedua kelas tersebut. Salah satu faktor yang diduga memengaruhi kondisi ini adalah adanya kemiripan pola bahasa antar ulasan prioritas Sedang dan Rendah, sehingga karakteristik kedua kelas menjadi kurang terpisah secara jelas. Selain itu, ketidakseimbangan distribusi data juga dapat memengaruhi kemampuan model dalam

mempelajari pola pada masing-masing kelas secara optimal. Meskipun demikian, hasil penelitian menunjukkan bahwa model memiliki potensi yang baik dalam mendukung proses identifikasi otomatis terhadap keluhan prioritas tinggi pada ulasan pengguna aplikasi e-commerce.

Temuan tersebut juga didukung oleh hasil *normalized confusion matrix* pada Gambar 2. Terlihat bahwa sebagian besar data dengan label Prioritas Sedang diprediksi sebagai Prioritas Tinggi sebesar 89,46%, sedangkan pada Prioritas Rendah mencapai 72,69%. Sebaliknya, model mampu mengklasifikasikan 97,75% data pada kelas Prioritas Tinggi dengan benar. Pola tersebut menunjukkan adanya kecenderungan model untuk memberikan prediksi ke kelas mayoritas, sehingga kemampuan dalam membedakan keluhan dengan tingkat urgensi Sedang dan Rendah masih belum optimal. Meskipun demikian, kemampuan model dalam mengenali keluhan dengan prioritas tinggi tetap memberikan manfaat yang signifikan karena memungkinkan sistem mengidentifikasi keluhan yang memerlukan penanganan segera.

3.2 Pembahasan

Hasil penelitian menunjukkan bahwa arsitektur *Hybrid Transformer-BiLSTM* mampu memberikan performa yang sangat baik pada tugas klasifikasi kategori keluhan pelanggan Tokopedia. Kombinasi IndoBERT dan BiLSTM terbukti efektif dalam memahami konteks semantik dan pola sekuensial pada teks ulasan berbahasa Indonesia.

Performa tinggi pada klasifikasi kategori menunjukkan bahwa setiap kategori memiliki karakteristik kata kunci yang cukup spesifik. Sebagai contoh, kategori Transaksi didominasi oleh kata-kata seperti “checkout”, “refund”, dan “pembayaran”, sedangkan kategori Teknis Aplikasi banyak mengandung kata seperti “error”, “crash”, dan “loading”. Pola semantik yang jelas tersebut memudahkan model dalam mempelajari hubungan antara teks ulasan dan label kategori.

Tingginya performa pada tugas klasifikasi kategori menunjukkan bahwa kombinasi representasi kontekstual dari IndoBERT dan kemampuan pemodelan sekuensial BiLSTM mampu saling melengkapi. IndoBERT berperan dalam memahami makna kata berdasarkan konteks kalimat, sedangkan BiLSTM membantu menangkap hubungan antar token dari dua arah sehingga informasi yang bersifat berurutan tetap dapat dipertahankan. Kombinasi kedua pendekatan tersebut memungkinkan model menghasilkan representasi fitur yang lebih kaya dibandingkan apabila hanya menggunakan salah satu arsitektur secara terpisah. Pendekatan multi-task learning memungkinkan model mempelajari representasi fitur yang digunakan secara bersama oleh dua tugas klasifikasi. Dengan adanya shared representation, informasi yang dipelajari dari klasifikasi kategori dapat turut mendukung proses pembelajaran pada klasifikasi prioritas. Pendekatan ini tidak hanya meningkatkan efisiensi pelatihan karena kedua tugas diproses dalam satu model, tetapi juga membantu model memanfaatkan keterkaitan antar tugas sehingga menghasilkan representasi yang lebih informatif.

Sebaliknya, klasifikasi prioritas menunjukkan performa yang lebih rendah dibandingkan klasifikasi kategori. Hal ini disebabkan oleh beberapa faktor. Pertama, distribusi data prioritas sangat tidak seimbang, di mana mayoritas data berada pada kelas Tinggi. Kedua, penentuan prioritas tidak hanya dipengaruhi oleh kata kunci, tetapi juga oleh konteks kalimat, tingkat keparahan, dan intensitas keluhan. Akibatnya, batas antara prioritas Sedang dan Rendah menjadi lebih sulit dipelajari model. Selain dipengaruhi oleh ketidakseimbangan distribusi data, proses penentuan tingkat prioritas juga memiliki kompleksitas yang lebih tinggi dibandingkan klasifikasi kategori. Hal ini disebabkan karena penentuan prioritas tidak hanya bergantung pada keberadaan kata kunci tertentu, tetapi juga mempertimbangkan konteks kalimat, tingkat keparahan masalah, serta implikasi yang ditimbulkan terhadap pengguna. Oleh karena itu, dua ulasan dengan topik yang serupa dapat memiliki tingkat prioritas yang berbeda bergantung pada konteks penyampaiannya. Kondisi tersebut menyebabkan model memerlukan pemahaman semantik yang lebih mendalam untuk dapat membedakan setiap tingkat prioritas secara konsisten. Selain itu, penggunaan *pseudo labeling* berbasis aturan memberikan kontribusi penting terhadap keberhasilan model. Pendekatan ini memungkinkan proses pelabelan dilakukan secara otomatis pada dataset berukuran besar tanpa anotasi manual. Hasil evaluasi menunjukkan bahwa label yang dihasilkan cukup konsisten, khususnya pada tugas klasifikasi kategori.

Rendahnya nilai macro F1-score terutama dipengaruhi oleh ketidakseimbangan distribusi kelas pada tugas klasifikasi prioritas. Berdasarkan confusion matrix pada Gambar 2, sebagian besar data Prioritas Sedang dan Rendah diprediksi sebagai Prioritas Tinggi, sehingga menghasilkan nilai recall yang rendah pada kedua kelas minoritas. Pola tersebut menunjukkan adanya bias model terhadap kelas mayoritas (majority class bias), meskipun kemampuan model dalam mengenali kelas Prioritas Tinggi tetap sangat baik. Kondisi ini umum dijumpai pada permasalahan klasifikasi dengan distribusi kelas yang tidak seimbang, karena fungsi optimasi cenderung lebih banyak dipengaruhi oleh kelas yang memiliki jumlah sampel lebih besar. Penelitian ini belum menerapkan teknik penanganan class imbalance seperti oversampling (misalnya SMOTE), focal loss, maupun cost-sensitive learning karena tujuan penelitian difokuskan pada evaluasi kemampuan dasar arsitektur Hybrid IndoBERT-BiLSTM berbasis multi-task learning terhadap distribusi data yang merepresentasikan kondisi nyata ulasan pelanggan Tokopedia. Penerapan teknik-teknik tersebut berpotensi meningkatkan kemampuan model dalam mengenali kelas minoritas, khususnya Prioritas Sedang dan Rendah, sehingga menjadi salah satu arah pengembangan yang akan dievaluasi pada penelitian selanjutnya. Dominasi prediksi menuju kelas Prioritas Tinggi menunjukkan bahwa nilai akurasi keseluruhan sebesar 80,04% belum sepenuhnya merepresentasikan kemampuan model pada seluruh kelas. Oleh karena itu, penelitian ini juga

menggunakan precision, recall, F1-score, macro F1-score, serta confusion matrix sebagai metrik evaluasi agar performa setiap kelas dapat dianalisis secara lebih komprehensif.

Penggunaan *pseudo labeling* juga memiliki keterbatasan karena kualitas label sangat bergantung pada aturan dan kata kunci yang digunakan. Kesalahan pada proses pelabelan otomatis dapat memengaruhi performa model, terutama pada kelas yang memiliki konteks ambigu atau tumpang tindih antar kategori. Secara keseluruhan, hasil penelitian menunjukkan bahwa pendekatan *multi-task learning* menggunakan *Hybrid Transformer-BiLSTM* mampu menjadi solusi yang efektif untuk klasifikasi keluhan pelanggan e-commerce. Model tidak hanya mampu mengidentifikasi jenis keluhan, tetapi juga menentukan tingkat prioritas penanganannya secara simultan. Dengan kemampuan tersebut, model berpotensi membantu perusahaan e-commerce dalam melakukan otomatisasi *routing* keluhan dan penentuan urgensi penanganan pelanggan secara lebih cepat dan efisien. Dari sisi implementasi, model yang diusulkan berpotensi mendukung sistem manajemen layanan pelanggan pada platform e-commerce. Informasi mengenai kategori dan prioritas keluhan dapat dimanfaatkan untuk melakukan proses *automatic complaint routing*, yaitu mengarahkan keluhan kepada unit penanganan yang sesuai sekaligus menentukan tingkat urgensi penanganannya. Dengan demikian, proses penanganan keluhan dapat dilakukan secara lebih cepat dan konsisten dibandingkan proses klasifikasi manual. Selain meningkatkan efisiensi proses klasifikasi, penerapan model ini juga berpotensi membantu perusahaan dalam menentukan prioritas alokasi sumber daya layanan pelanggan. Keluhan dengan prioritas tinggi dapat segera diteruskan kepada tim yang berwenang, sedangkan keluhan dengan prioritas lebih rendah dapat diproses sesuai antrean layanan. Dengan demikian, waktu respons terhadap keluhan yang bersifat kritis dapat dipersingkat sehingga berpotensi meningkatkan kualitas pelayanan dan kepuasan pelanggan.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa arsitektur Hybrid IndoBERT–BiLSTM berbasis multi-task learning mampu diterapkan untuk melakukan klasifikasi kategori dan prioritas keluhan pelanggan Tokopedia secara simultan pada ulasan berbahasa Indonesia. Kombinasi IndoBERT sebagai pembentuk representasi kontekstual dan BiLSTM sebagai pemodel hubungan sekuensial menghasilkan performa yang sangat baik pada klasifikasi kategori, sekaligus menunjukkan bahwa pembelajaran bersama (shared representation) mampu dimanfaatkan untuk menyelesaikan dua tugas yang saling berkaitan dalam satu model. Penelitian ini memberikan dua kontribusi utama. Pertama, penelitian ini menunjukkan bahwa arsitektur Hybrid IndoBERT–BiLSTM berbasis Multi-Task Learning dapat diterapkan untuk mengklasifikasikan kategori dan prioritas keluhan pelanggan secara simultan pada ulasan e-commerce berbahasa Indonesia. Kedua, penelitian ini menghasilkan dataset keluhan Tokopedia yang telah melalui proses *pseudo labeling* berbasis aturan sehingga dapat dimanfaatkan sebagai referensi bagi penelitian lanjutan di bidang Natural Language Processing (NLP) dan manajemen layanan pelanggan digital. Meskipun demikian, penelitian ini masih memiliki beberapa keterbatasan. Kualitas label yang dihasilkan masih bergantung pada aturan dan kata kunci yang digunakan, sedangkan ketidakseimbangan distribusi data pada label prioritas menyebabkan model lebih dominan mengenali kelas mayoritas sehingga performa pada kelas minoritas masih belum optimal. Oleh karena itu, penelitian selanjutnya dapat mengembangkan mekanisme pelabelan yang mengombinasikan *pseudo labeling* dengan validasi anotator (human-in-the-loop), serta menerapkan strategi penanganan imbalanced dataset seperti class weighting, focal loss, atau teknik resampling untuk meningkatkan kemampuan model dalam mengenali kelas minoritas. Selain itu, penambahan informasi kontekstual, seperti riwayat respons layanan pelanggan, waktu penyelesaian keluhan, maupun metadata transaksi, berpotensi menghasilkan representasi yang lebih kaya sehingga sistem klasifikasi dan prioritisasi keluhan dapat mendukung proses pengambilan keputusan pada layanan pelanggan secara lebih akurat dan efektif.

5. PERNYATAAN PENULIS

5.1 Kontribusi Penulis

Konseptualisasi: R. H. S., dan A. M. S.;

Metodologi: R. H. S., dan A. M. S.;

Perangkat Lunak: R. H. S., dan A. M. S.;

Validasi: R. H. S., dan A. M. S.;

Analisis Formal: R. H. S., dan A. M. S.;

Investigasi: R. H. S., dan A. M. S.;

Sumber Daya: R. H. S., dan A. M. S.;

Kurasi Data: R. H. S.;

Penulisan Draf Awal: R. H. S., dan A. M. S.;

Penulisan, Peninjauan, dan Penyuntingan: R. H. S., dan A. M. S.;

Visualisasi: R. H. S., dan A. M. S.;

Seluruh penulis telah membaca dan menyetujui versi manuskrip yang telah diterbitkan.

5.2 Ketersediaan Data

Data yang disajikan dalam penelitian ini tersedia berdasarkan permintaan kepada penulis korespondensi.

5.3 Pendanaan

Penelitian ini tidak menerima pendanaan eksternal.

5.4 Pernyataan Dewan Peninjau Institusional

Tidak Berlaku (Not applicable).

5.5 Pernyataan Persetujuan Berdasarkan Informasi

Tidak Berlaku (Not applicable).

5.6 Pernyataan Konflik Kepentingan

Para penulis menyatakan bahwa mereka tidak memiliki kepentingan keuangan yang bersaing atau hubungan pribadi yang diketahui yang dapat dianggap memengaruhi pekerjaan yang dilaporkan dalam artikel ini.

5.7 Penggunaan AI Generatif dan Teknologi Berbantuan AI dalam Proses Penulisan

Para penulis menggunakan AI Generatif untuk meningkatkan kejelasan penulisan dalam artikel ini. Para penulis telah meninjau dan menyunting konten yang dihasilkan dengan bantuan AI serta bertanggung jawab penuh atas versi akhir publikasi.

REFERENCES

- [1] K. Kusumahadi, W. Sastika, and F. H. Hanifa, "E-ServiceQuality Model and Kano Model: An Integration Improvement on Service Quality in E-Commerce Services," *International Journal of Entrepreneurship, Business and Creative Economy*, vol. 2, no. 1, pp. 67–80, Jan. 2022, doi: 10.31098/ijebce.v2i1.738.
- [2] R. S. Handayani, G. Alysah Diva, E. N. Hermansyah, and A. Siswati, "IKN: Jurnal Informatika dan Kesehatan Transformasi Digital E-Commerce di Era Post-Pandemic: Analisis Tren dan Dampak pada Perilaku Konsumen Indonesia," *IKN: Jurnal Informatika dan Kesehatan*, vol. 2, no. 2, Aug. 2025, doi: 10.35473/ikn.v2i2.3704.
- [3] R. P. Putri and M. I. Fasa, "Transformasi Digital: Dampak E-Commerce terhadap Perkembangan E-Business di Indonesia," *Journal of Social Science and Multidisciplinary Analysis*, vol. 2, no. 2, Apr. 2025, doi: 10.59827/jossama.v2i2.51.
- [4] N. Jauharin Insi'ah, R. Itsnainiyah, Z. Ghina, and M. K. Milad, "Analisis Risiko Keluhan Pelanggan dan Upaya Mitigasi di PT. PLN (Persero) UP3 Gresik Customer Complaint Risk Analysis and Mitigation Efforts at PT. PLN (Persero) UP3 Gresik," *Jurnal Kolaboratif Sains*, vol. 8, no. 5, pp. 2175–2187, 2025, doi: 10.56338/jks.v8i5.7398.
- [5] Rahmi Yuliana and Rinwanti Rinwanti, "Pengaruh ServQual Terhadap EWOM Melalui Trust dan E-Satisfaction Online Merchant," *Jurnal Visi Manajemen*, vol. 11, no. 2, pp. 26–40, Apr. 2025, doi: 10.56910/jvm.v11i2.616.
- [6] S. Sandika, W. Hardinata, and N. Y. Maleha, "Analisis Pengaruh Kualitas Pelayanan, Persepsi Harga Dan Kepuasan Terhadap Loyalitas Konsumen (Studi Kasus Pada E-Commerce Shopee)," *Jurnal Ekonomi Bisnis, Manajemen dan Akuntansi (JEBMAK)*, vol. 4, no. 1, Mar. 2025, doi: 10.61930/jebmak.v4i1.1119.
- [7] A. R. Mahendra and Mugiono, "Pengaruh E-Service Quality, E-Word Of Mouth, Dan E-Customer Trust Terhadap Repurchase Intention Pengguna E-Commerce," *Jurnal Manajemen Pemasaran dan Perilaku Konsumen*, vol. 1, no. 4, pp. 476–486, Oct. 2022, doi: 10.21776/jmppk.2022.01.4.07.
- [8] Elia Cahyani, Tazkiyyah Nurrohman, Cindy Firstiananda Deka, and Mohamad Zein Saleh, "The Role of Service Quality in Building Customer Satisfaction," *Journal of Management and Creative Business*, vol. 2, no. 4, pp. 192–205, Oct. 2024, doi: 10.30640/jmcbs.v2i4.3300.
- [9] Fine Brilian and D. P. Novalita, "Strategi Penanganan Keluhan Wisatawan Dalam Menjaga Citra Dan Loyalitas Di Hotel Grafika Cikole, Lembang," *Jurnal Media Akademik (JMA)*, vol. 3, no. 6, pp. 3031–5220, Jun. 2025, doi: 10.62281/v3i6.2041.
- [10] D. Z. Robert, A. B. Setiawan, and D. W. Widodo, "Implementasi Metode Support Vector Machine untuk Otomatis Pengaduan Publik di Kabupaten Trenggalek," *INOTEK*, vol. 9, no. 2, p. 1617, 2025, doi: 10.29407/bw7jd239.
- [11] S. Amanda Salsabila, B. Priyatna, A. Hananto, and Tukino, "Komparasi Kinerja Model Naive Bayes, SVM, dan BERT dalam Klasifikasi Sentimen Ulasan Pada Aplikasi YUMMY," *STORAGE*, vol. 4, no. 2, May 2025, doi: 10.55123/storage.v4i2.5120.
- [12] A. R. Aditama and A. F. Wicaksono, "Classification of customer complaints on social media for e-commerce in Indonesia," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 15, no. 3, p. 2977, Jun. 2025, doi: 10.11591/ijece.v15i3.pp2977-2985.
- [13] Athira, M. Adith, and D. Gupta, "Effective Complaint Detection in Financial Services through Complaint, Severity, Emotion and Sentiment Analysis," in *Procedia Computer Science*, Elsevier B.V., 2025, pp. 2220–2231. doi: 10.1016/j.procs.2025.04.472.
- [14] V. D. Setiawan and D. U. Iswawigra, "Sentiment Analysis to Evaluate Public Service Perception among Surakarta City Resident Using the BiLSTM Model," *JITE (Journal of Informatics and Telecommunication Engineering)*, vol. 9, no. 1, Jul. 2025, doi: 10.31289/jite.v9i1.15498.
- [15] M. Ridwan Usmana and A. Rais Ruli, "Analisis Sentimen Ulasan Pengguna Terhadap Kualitas Layanan Transportasi Online Pada Aplikasi Kai Access Menggunakan Metode Deep Learning Dengan Algoritma Lstm, Gru, Dan Cnn," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 5, Oct. 2025, doi: 10.36040/jati.v9i5.14744.
- [16] Alfando and R. Hayami, "Klasifikasi Teks Berita Berbahasa Indonesia Menggunakan Machine Learning Dan Deep Learning: Studi Literatur," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 1, Feb. 2023, doi: 10.36040/jati.v7i1.6486.

- [17] B. R. Bhamare and J. Prabhu, “A multilabel classifier for text classification and enhanced BERT system,” *Revue d’Intelligence Artificielle*, vol. 35, no. 2, pp. 167–176, Apr. 2021, doi: 10.18280/ria.350209.
- [18] H. Ahmadian, T. F. Abidin, H. Riza, and K. Muchtar, “Hybrid Models for Emotion Classification and Sentiment Analysis in Indonesian Language,” *Applied Computational Intelligence and Soft Computing*, vol. 2024, 2024, doi: 10.1155/2024/2826773.
- [19] Arif Fitra Setyawan, Amelia Devi Putri Ariyanto, Fari Katul Fikriah, and Rozaq Isnaini Nugraha, “Analisis Sentimen Ulasan iPhone di Amazon Menggunakan Model Deep Learning BERT Berbasis Transformer,” *Elkom: Jurnal Elektronika dan Komputer*, vol. 17, no. 2, pp. 447–452, Dec. 2024, doi: 10.51903/elkom.v17i2.2150.
- [20] A. S. Rizkia, W. Wufiron, and F. F. Roji, “Analisis Sentimen Coretax: Perbandingan Pelabelan Data Manual, Transformers-Based, dan Lexicon-Based pada Performa IndoBERT,” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 5, no. 3, Jul. 2025, doi: 10.57152/malcom.v5i3.2151.
- [21] F. Moch. Habibi, R. Gelar Guntara, and A. Nuryadin, “Komparasi Model IndoBERT dan IndoGPT untuk Analisis Sentimen pada Produk E-commerce,” *Jurnal Algoritma*, vol. 22, no. 2, Dec. 2025, doi: 10.33364/algoritma/v.22-2.3070.
- [22] A. Zevana and D. Riana, “Text Classification Using Indobert Fine-Tuning Modeling With Convolutional Neural Network And Bi-Lstm,” *Jurnal Teknik Informatika (Jutif)*, vol. 4, no. 6, pp. 1605–1610, Jan. 2024, doi: 10.52436/1.jutif.2023.4.6.1650.
- [23] F. K. Ihtada, R. Alfianita, and O. Q. Aziz, “Aspect-based Multilabel Classification of E-commerce Reviews Using Fine-tuned IndoBERT,” *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, Jan. 2025, doi: 10.22219/kinetik.v10i1.2088.
- [24] M. M. Hossain, M. S. Hossain, M. Safran, S. Alfarhood, M. Alfarhood, and M. F. Mridha, “A Hybrid Attention-Based Transformer Model for Arabic News Classification Using Text Embedding and Deep Learning,” *IEEE Access*, vol. 12, pp. 198046–198066, 2024, doi: 10.1109/ACCESS.2024.3522061.
- [25] J. Cui, F. Zhang, and Y. Zong, “An Attention-Guided Bert-BiLSTM Model for Chinese Sentiment Classification,” *Association for Computing Machinery*, pp. 100–108, Jan. 2026, doi: 10.1145/3779475.3779492.
- [26] S. Chen, Y. Zhang, and Q. Yang, “Multi-Task Learning in Natural Language Processing: An Overview,” *ACM Comput. Surv.*, vol. 56, no. 12, Jul. 2024, doi: 10.1145/3663363.
- [27] Z. Zhang, W. Yu, M. Yu, Z. Guo, and M. Jiang, “A Survey of Multi-task Learning in Natural Language Processing: Regarding Task Relatedness and Training Methods,” *ACL Anthology*, pp. 943–956, May 2023, doi: 10.18653/v1/2023.acl-main.66.
- [28] J. Ni, Z. Jin, Q. Wang, M. Sachan, E. Zürich, and M. Leippold, “When Does Aggregating Multiple Skills with Multi-Task Learning Work? A Case Study in Financial NLP,” *ACL Anthology*, vol. 1, pp. 7465–7488, Jul. 2023, doi: 10.18653/v1/2023.acl-long.412.
- [29] K. Taneja, J. Vashishtha, and S. Ratnoo, “Transformer Based Unsupervised Learning Approach for Imbalanced Text Sentiment Analysis of E-Commerce Reviews,” in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 2318–2331. doi: 10.1016/j.procs.2024.04.220.