

Aspect-Category Sentiment Analysis for Knowledge Discovery in ShopeePay App Reviews Using Fine-Tuned IndoBERT

Nabila Ramadhani¹, Ken Ditha Tania^{2*}, Meylani Utari³

^{1,2,3} Sistem Informasi, Universitas Sriwijaya

nabila.r086@gmail.com¹, kenya.tania@gmail.com², meylani@ilkom.unsri.ac.id³

Article Info

Article history:

Received 2026-07-27

Revised 2026-09-11

Accepted 2026-10-15

Keyword:

ACSA,
Aspect Category Sentiment
Analysis,
IndoBERT,
Knowledge Discovery,
ShopeePay.

ABSTRACT

The rapid growth of e-wallet adoption in Indonesia has generated large volumes of user reviews on Google Play Store, yet document-level sentiment analysis fails to capture that a single review may express different sentiments toward different service aspects. This study applies Aspect-Category Sentiment Analysis (ACSA), a sub-task of Aspect-Based Sentiment Analysis (ABSA) combining Aspect Category Detection (ACD) and Sentiment Classification (ASC), to classify sentiment across five predefined aspects of ShopeePay reviews: Security, Ease of Use/UI, Performance, Features, and Customer Service. From 42,145 scraped reviews, 22,756 unique reviews remained after preprocessing, of which 12,525 (55.0%) contained at least one relevant aspect. Labels were generated using Gemini 3.5 Flash-Lite and validated against 1,500 manually annotated samples via Cohen's Kappa (ACD: $\kappa = 0.744$; sentiment: $\kappa = 0.903$), with ACD further evaluated against the gold-standard (macro-F1 = 0.794). Using 15,986 binary-labeled instances, IndoBERT was fine-tuned via stratified 5-fold cross-validation and compared against SVM, Naive Bayes, and IndoRoBERTa. IndoBERT consistently outperformed all baselines across the five aspects (F1 = 0.970–0.984), with the largest gap for Ease of Use/UI (0.970 vs. 0.788), and results were confirmed on a human-annotated subset (F1 = 0.971–1.000). Knowledge discovery on the model's predictions showed Features and Performance were the most frequently complained-about aspects (3,035 and 1,455 negative reviews, respectively), suggesting these areas require urgent technical attention, while Ease of Use/UI received the highest volume of positive sentiment (3,819 reviews), indicating its interface design should be preserved rather than redesigned, offering developers a prioritized, data-driven basis for improvement.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

I. PENDAHULUAN

Perkembangan teknologi digital telah mendorong transformasi sistem pembayaran menuju transaksi non-tunai melalui layanan dompet digital (*e-wallet*). Sebagai salah satu inovasi dalam sistem pembayaran digital, *e-wallet* berkembang pesat seiring meningkatnya aktivitas transaksi masyarakat di dunia nyata maupun dunia maya [1]. Pertumbuhan *e-wallet* yang didukung oleh perkembangan *e-commerce* dan pembayaran seluler turut mempercepat terwujudnya masyarakat tanpa uang tunai [2]. Di Indonesia, ShopeePay menjadi salah satu *e-wallet* dengan tingkat adopsi yang tinggi. Berdasarkan survei [3], ShopeePay mendominasi

penggunaan *e-wallet* pada *merchant* dengan pangsa penggunaan sebesar 57%, melampaui GoPay, DANA, dan OVO. Tingginya penggunaan ShopeePay menghasilkan banyak ulasan pengguna pada Google Play Store yang memuat opini terhadap berbagai aspek layanan [4]. Ulasan tersebut menjadi sumber informasi yang penting untuk mengevaluasi kualitas layanan serta memahami kebutuhan pengguna.

Ulasan pengguna dapat dimanfaatkan sebagai bahan evaluasi bagi pengembang untuk mengetahui tingkat kepuasan pengguna serta meningkatkan kualitas layanan aplikasi [5]. Namun, banyaknya ulasan menyebabkan analisis secara manual menjadi tidak efisien, memerlukan waktu yang

lama, dan sulit dilakukan secara konsisten [6]. Selain itu, karakteristik ulasan pada Google Play Store yang tidak terstruktur, bersifat informal, mengandung singkatan, serta memiliki variasi linguistik yang tinggi juga menyulitkan proses analisis menggunakan pendekatan konvensional [7].

Salah satu pendekatan untuk mengatasi permasalahan tersebut adalah analisis sentimen, yaitu teknik yang memanfaatkan *text mining* dan *Natural Language Processing* (NLP) untuk mengklasifikasikan ulasan ke dalam kategori sentimen positif, negatif, atau netral [8]. Meskipun mampu mengotomatisasi proses analisis opini pengguna, analisis sentimen pada tingkat dokumen umumnya hanya menghasilkan satu label sentimen untuk keseluruhan ulasan. Padahal, satu ulasan dapat memuat beberapa aspek dengan polaritas sentimen yang berbeda sehingga pendekatan tersebut belum mampu mengidentifikasi sentimen pada setiap aspek layanan secara terperinci [9].

Untuk mengatasi keterbatasan tersebut, dikembangkan pendekatan *Aspect-Based Sentiment Analysis* (ABSA). ABSA merupakan bentuk analisis sentimen yang lebih terperinci (*fine-grained sentiment analysis*) karena tidak hanya mengidentifikasi polaritas sentimen suatu ulasan, tetapi juga mengenali elemen-elemen sentimen yang terkandung di dalamnya, seperti aspek yang dibahas, kategori aspek, istilah opini, serta polaritas sentimen pada setiap aspek [9]. Salah satu sub-task dalam kerangka ABSA adalah *Aspect-Category Sentiment Analysis* (ACSA), yaitu pendekatan yang menggabungkan *Aspect Category Detection* (ACD) proses mengidentifikasi kategori aspek yang dibahas dalam suatu teks berdasarkan himpunan kategori yang telah ditetapkan sebelumnya (*predefined categories*) dengan *Aspect Category Sentiment Classification* (ASC), yaitu proses mengklasifikasikan polaritas sentimen terhadap kategori aspek tersebut [10]. Dengan demikian, ABSA mampu menghasilkan informasi yang lebih rinci mengenai sentimen pada setiap aspek layanan sehingga dapat membantu pengembang menentukan prioritas perbaikan aplikasi secara lebih tepat sasaran [5], [6]. Pendekatan ini telah banyak diterapkan pada berbagai domain, seperti *e-commerce*, keuangan, kesehatan, pendidikan, dan media sosial [9].

Hasil klasifikasi ABSA juga dapat mendukung *Knowledge Discovery* yaitu proses menemukan pengetahuan baru yang valid dan bermanfaat dari data yang sebelumnya tidak diketahui secara eksplisit [11]. Hasil klasifikasi ABSA terhadap ulasan pengguna tidak hanya berhenti pada pelabelan kategori aspek dan polaritas sentimen, tetapi diinterpretasikan lebih lanjut melalui analisis distribusi dan frekuensi sentimen pada setiap kategori aspek untuk mengungkap pengetahuan mengenai pola persepsi pengguna seperti aspek yang paling sering dikeluhkan, aspek yang paling diapresiasi, serta kecenderungan sentimen pada setiap kategori aspek secara keseluruhan. Interpretasi ini diperkaya melalui ekstraksi kata kunci berbasis TF-IDF dan topic modeling menggunakan Latent Dirichlet Allocation (LDA) untuk mengungkap pola permasalahan spesifik di balik setiap kategori aspek, serta didukung oleh sampel ulasan

representatif sebagai bukti kualitatif yang memperjelas temuan kuantitatif tersebut. Pengetahuan yang dihasilkan dari proses tersebut selanjutnya dapat dimanfaatkan sebagai dasar pengambilan keputusan bagi pengembang ShopeePay dalam menentukan prioritas perbaikan maupun mempertahankan aspek layanan yang telah dinilai baik oleh pengguna. Kerangka *knowledge discovery* ini telah diterapkan pada ulasan aplikasi layanan keuangan digital untuk mengungkap wawasan mengenai persepsi keamanan pengguna [12], serta pada domain layanan keuangan digital lainnya untuk mendukung pengambilan keputusan bisnis [13], [14].

Seiring berkembangnya metode berbasis *deep learning*, model berbasis Transformer menjadi salah satu pendekatan yang banyak dimanfaatkan dalam berbagai tugas NLP. Untuk bahasa Indonesia, IndoBERT merupakan model *pre-trained Transformer* yang dirancang untuk memahami karakteristik linguistik bahasa Indonesia sehingga mampu menghasilkan representasi kontekstual yang lebih baik dibandingkan algoritma pembelajaran mesin konvensional [15], [16]. Berbagai penelitian menunjukkan bahwa IndoBERT lebih adaptif terhadap kompleksitas bahasa Indonesia dibandingkan algoritma konvensional seperti SVM dan NB yang cenderung kesulitan menangani slang dan ekspresi informal pengguna Indonesia, sebagaimana dilaporkan pada studi-studi terdahulu [7], serta secara konsisten menghasilkan performa yang lebih tinggi pada tugas analisis sentimen berbahasa Indonesia [17], [18], [19]. Selain itu, penelitian [6] menunjukkan bahwa model IndoBERT yang di-*fine-tuning* memperoleh performa terbaik pada tugas ABSA. Meskipun demikian, hasil tersebut diperoleh pada domain yang berbeda sehingga masih diperlukan evaluasi untuk mengetahui konsistensi performa Fine-Tuned IndoBERT pada ulasan aplikasi ShopeePay serta membandingkannya dengan algoritma konvensional.

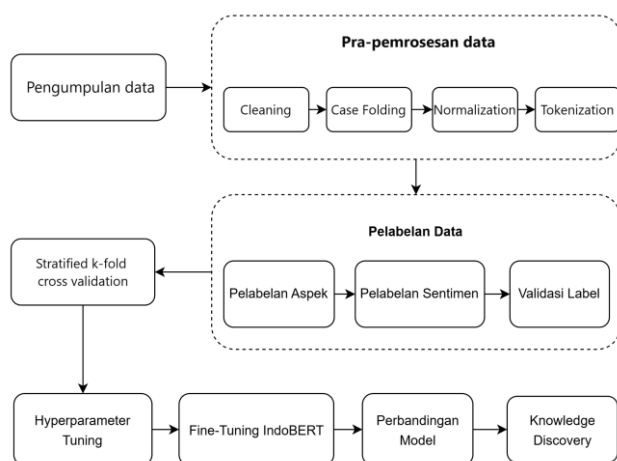
Meskipun berbagai penelitian telah menerapkan analisis sentimen maupun ABSA pada berbagai domain dan menunjukkan potensi IndoBERT dalam meningkatkan performa klasifikasi, penelitian mengenai ShopeePay masih berfokus pada faktor-faktor yang memengaruhi niat atau perilaku penggunaan aplikasi [20], sedangkan kajian yang menganalisis opini pengguna berdasarkan setiap aspek layanan secara *fine-grained* masih terbatas. Di sisi lain, penelitian ABSA berbahasa Indonesia yang telah memanfaatkan model berbasis *Transformer* sejauh ini umumnya diterapkan pada domain di luar layanan *e-wallet*, seperti transportasi [21], kesehatan mental [22], layanan rumah sakit [23] dan belum banyak yang secara khusus mengevaluasi konsistensi performa IndoBERT pada karakteristik ulasan *e-wallet* yang cenderung singkat, informal, dan padat istilah transaksi. Kondisi ini menunjukkan adanya celah penelitian yang perlu dikaji lebih lanjut, baik dari sisi domain aplikasi maupun dari sisi perbandingan performa model. Selain itu, penerapan ABSA pada ulasan aplikasi berbahasa Indonesia, khususnya menggunakan Fine-Tuned IndoBERT, juga masih relatif terbatas [6], [17]. Berdasarkan celah tersebut, penelitian ini

memberikan kontribusi ilmiah dengan menerapkan ACSA menggunakan Fine-Tuned IndoBERT pada domain layanan *e-wallet* yang belum dikaji penelitian Transformer Indonesia sebelumnya, sekaligus mengevaluasi performanya secara komparatif terhadap algoritma SVM, NB dan IndoRoBERTa pada dataset dan skema evaluasi yang sama, yang belum banyak dilakukan pada penelitian ABSA Transformer Indonesia terdahulu.

Berdasarkan kontribusi tersebut, penelitian ini secara khusus bertujuan mengidentifikasi sentimen pada setiap aspek layanan ShopeePay melalui pendekatan ACSA, mengevaluasi konsistensi performa *Fine-Tuned IndoBERT* dibandingkan SVM, NB dan IndoRoBERTa, serta menerapkan tahap *knowledge discovery* pada hasil klasifikasi model terbaik untuk menggali wawasan mengenai aspek yang paling banyak dikeluhkan maupun diapresiasi pengguna. Hasil penelitian diharapkan dapat memberikan kontribusi akademis berupa evaluasi performa model pada domain ulasan aplikasi *e-wallet* serta memberikan kontribusi praktis berupa rekomendasi bagi pengembang ShopeePay mengenai aspek-aspek layanan yang perlu dipertahankan maupun ditingkatkan berdasarkan opini pengguna.

II. METODE

Penelitian ini menerapkan tahapan yang tersusun secara sistematis dan berurutan dari tahap awal hingga tahap akhir, mengikuti kerangka *knowledge discovery* yang mencakup proses seleksi data, praprosesing, transformasi, penambahan data, hingga evaluasi hasil [11], sebagaimana juga diterapkan pada penelitian sejenis di domain aplikasi keuangan digital [12], [13]. Gambar 1 menampilkan alur penelitian dalam bentuk diagram alur (*flowchart*).



Gambar 1. Alur Penelitian

A. Pengumpulan Data

Data penelitian berfokus pada ulasan pengguna ShopeePay di Google Play Store, sehingga belum mencakup ulasan dari App Store atau media sosial. Google Play Store dipilih sebagai sumber data karena menyediakan jumlah ulasan yang jauh lebih banyak dibandingkan App Store, sehingga lebih

representatif untuk dianalisis. Pengumpulan data dilakukan dengan teknik web scraping menggunakan pustaka *google-play-scraper* berbasis Python. Teknik ini umum digunakan dalam penelitian ABSA berbahasa Indonesia karena memungkinkan pengambilan data ulasan pengguna secara otomatis, lengkap dengan metadata rating dan tanggal, tanpa memerlukan akses API resmi dari penyedia aplikasi [24]. Proses pengambilan data dilakukan dalam rentang waktu Januari hingga Juni 2026 untuk memperoleh data ulasan terkini yang mencerminkan pengalaman dan persepsi pengguna secara aktual.

B. Pra-pemrosesan Data

Pra-pemrosesan data merupakan tahap penting dalam alur kerja *Natural Language Processing* (NLP) karena kualitas hasil pemrosesan pada tahap ini mempengaruhi akurasi model pada tahap selanjutnya. Penulis menerapkan empat proses berurutan pada tahap ini, yaitu *cleaning*, *case folding*, *normalization*, dan *tokenization*.

1) *Cleaning*: Teks dibersihkan dari elemen yang tidak relevan pada tahap *cleaning* ini, meliputi tautan (URL), tanda baca, angka, karakter khusus, emoji, dan spasi berlebih, sementara kata negasi tetap dipertahankan agar makna sentimen kalimat tidak berubah. Data duplikat dan data spam turut dihapus untuk menjaga kualitas dataset. Data duplikat merupakan ulasan dengan isi teks identik atau sangat mirip akibat pengiriman berulang oleh pengguna, sedangkan data spam mencakup ulasan tidak relevan seperti promosi, tautan iklan, atau teks tanpa makna. Identifikasi data spam dilakukan menggunakan aturan berbasis panjang teks minimum, sehingga hanya ulasan yang relevan dan bermakna yang digunakan pada tahap pemrosesan selanjutnya.

2) *Case Folding*: Seluruh huruf pada teks diseragamkan menjadi huruf kecil (*lowercase*) pada tahap *case folding* ini. Proses ini mengurangi variasi representasi kata yang sebenarnya memiliki makna sama, misalnya kata "Bagus" dan "bagus". Kedua bentuk kata tersebut kemudian diperlakukan sebagai satu token yang identik oleh model.

3) *Normalization*: Kata tidak baku, singkatan, dan kata gaul (*slang*) diubah menjadi kata baku Bahasa Indonesia pada tahap normalisasi ini, misalnya kata "gak" menjadi "tidak" dan "bgt" menjadi "banget".

4) *Tokenization*: Pada tahap tokenisasi, teks dipecah menjadi satuan kata (*token*) menggunakan IndoBERT Tokenizer dari pustaka *HuggingFace Transformers*. Penelitian ini menggunakan varian *indobert-base-p2* yang menerapkan algoritma *WordPiece* untuk memecah kata menjadi *subword* apabila kata tersebut tidak terdapat dalam kosakata (*vocabulary*) model. Selanjutnya, setiap token dikonversi menjadi representasi numerik berupa input ID yang digunakan sebagai masukan bagi model IndoBERT dalam proses klasifikasi sentimen.

C. Pelabelan Data

Penetapan kategori aspek dalam penelitian ini didasarkan pada tinjauan sistematis terhadap literatur domain *mobile payment* dan *e-wallet*. Empat faktor kualitas utama dari ulasan aplikasi pembayaran *mobile* meliputi *payment convenience*, *security*, *customer service*, dan *app design* [4]. Dimensi *e-wallet* yang paling berpengaruh terhadap kepuasan pengguna meliputi *service quality*, *ease of use*, *usefulness*, *cost*, dan *security* [2]. Dimensi *privacy and security* serta *compatibility* juga merupakan atribut kritis pada aplikasi *mobile payment* [25], dan konsistensi aspek-aspek ini juga dikonfirmasi oleh *systematic review* adopsi *mobile payment* [26].

Dengan menggabungkan temuan dari literatur tersebut, penelitian ini menetapkan Keamanan, Kemudahan UI, Performa, Fitur, dan Layanan Pelanggan sebagai kategori aspek yang relevan secara domain untuk analisis ulasan ShopeePay. Aspek Keamanan dipilih karena *security* dan *privacy* terbukti memengaruhi kepuasan pengguna, bahkan lemahnya keamanan yang dirasakan menjadi salah satu faktor pendorong pengguna berhenti menggunakan aplikasi pembayaran *mobile* secara permanen [27]. Aspek Kemudahan UI ditetapkan karena *app design* dan *ease of use* tidak hanya memengaruhi kenyamanan interaksi jangka pendek, tetapi juga terbukti signifikan memengaruhi niat pengguna untuk terus menggunakan aplikasi dalam jangka panjang [25]. Aspek Performa dipilih karena *payment convenience* dan *compatibility* mencerminkan kelancaran transaksi pada berbagai kondisi perangkat dan jaringan, serta kualitas performa aplikasi terbukti memengaruhi *continuance intention* pengguna melalui peningkatan kepercayaan dan kepuasan [25], [28]. Aspek Fitur ditetapkan karena *usefulness* secara signifikan memengaruhi *intention to continue use* dan berperan sebagai mediator penting dalam membentuk kepuasan pengguna *e-wallet* [29]. Terakhir, aspek Layanan Pelanggan dipilih karena *service quality* secara signifikan memengaruhi loyalitas pelanggan pada layanan *mobile e-wallet* dalam jangka panjang [2].

Setelah kelima aspek ditetapkan berdasarkan hasil sintesis literatur tersebut, Penulis menyusun deskripsi dan aturan pembeda (*disambiguation rules*) untuk masing-masing aspek sebagai acuan bagi model bahasa besar (LLM) dalam mengidentifikasi aspek pada setiap ulasan. Deskripsi ini disusun secara iteratif, dimulai dari rumusan konseptual berdasarkan literatur, lalu diuji pada sampel ulasan berskala kecil untuk mengidentifikasi pola ambiguitas antar aspek yang berulang (misalnya Performa vs Kemudahan UI pada kasus verifikasi gagal, dan Fitur vs Performa pada kasus transaksi tidak berjalan), sebelum dirumuskan menjadi aturan prioritas eksplisit beserta contoh berlabel (*few-shot examples*) dan direvisi hingga siap diterapkan pada keseluruhan data. Deskripsi aspek yang digunakan dalam penelitian ini disajikan pada Tabel 1.

TABEL I
DESKRIPSI DAN ATURAN PEMBEDA ASPEK ULASAN

Aspek	Deskripsi
Keamanan	PIN, OTP, verifikasi wajah (dari sisi keamanan akses, bukan kegagalan teknisnya), akun diretas/dibobol, saldo hilang tanpa transaksi jelas, penipuan oleh pihak ketiga, data pribadi bocor, akun disalahgunakan orang lain
Kemudahan UI	Kemudahan atau kesulitan memakai aplikasi dari sisi desain/alur: tampilan membingungkan, proses login/daftar yang berbelit, langkah verifikasi berulang akibat alur/proses. Fokus pada pengalaman penggunaan (UX), bukan pada aplikasi yang <i>error/bug</i>
Performa	Aplikasi crash/force close/hang/error sistem, transaksi gagal karena sistem/server, loading lama, kecepatan aplikasi (cepat/lambat/lemot), log out sendiri, saldo tidak berubah akibat bug sistem
Fitur	Fungsi atau ketersediaan fitur spesifik: ShopeePayLater, QRIS, transfer, top up, cashback/promo, ShopeeFood, termasuk fitur dibekukan/ditolak/limit selama bukan disebabkan bug aplikasi
Layanan Pelanggan	Respons/sikap customer service, kecepatan penanganan komplain, proses klaim/refund yang melibatkan CS, termasuk penagihan yang terus menghubungi meski tagihan sudah lunas.

Selain deskripsi di atas, disusun pula seperangkat aturan prioritas untuk kasus-kasus ambigu yang sering muncul di dataset (misalnya membedakan kegagalan verifikasi akibat sistem versus akibat alur proses, serta membedakan kegagalan fungsi fitur versus bug sistem), beserta beberapa contoh berlabel (*few-shot examples*) yang disertakan dalam prompt untuk menjaga konsistensi keputusan model.

Proses pelabelan aspek dan sentimen pada penelitian ini dilakukan secara langsung menggunakan LLM (Gemini), yaitu model Gemini 3.5 Flash-Lite. Penulis memilih varian ini karena berstatus General Availability serta dirancang khusus untuk kebutuhan otomatisasi bervolume tinggi dengan latensi rendah dan biaya rendah, sehingga sesuai untuk kebutuhan pelabelan ulasan pada penelitian ini. Setiap ulasan dikirim ke model secara berkelompok bersama deskripsi aspek, aturan pembeda, dan contoh *few-shot*, dan model mengembalikan keluaran terstruktur yang memuat status keberadaan tiap aspek beserta sentimennya.

Satu ulasan dapat memuat lebih dari satu aspek layanan sekaligus dengan polaritas sentimen yang berbeda-beda antaraspek. Sebagai ilustrasi, kalimat "Makanannya enak, tapi pelayanannya buruk." memuat aspek makanan bersentimen positif dan aspek pelayanan bersentimen negatif [30]. Karakteristik inilah yang membuat skema keluaran bersifat multilabel, sehingga ulasan yang tidak membahas aspek manapun tetap dipertahankan sebagai kelas "tanpa aspek" yang sah, bukan dibuang dari dataset.

Label aspek dan sentimen awal ini kemudian divalidasi menggunakan Cohen's Kappa untuk mengukur kesepakatan

antara anotasi manual dan label otomatis tersebut [31]. Validasi diterapkan pada 1.500 data sampel yang dipilih melalui *stratified sampling* berdasarkan jumlah aspek yang terdeteksi oleh Gemini pada tiap ulasan (0 hingga 5 aspek), untuk memastikan proporsi ulasan sederhana maupun multi-aspek terwakili secara seimbang, kemudian dianotasi secara *blind* untuk menghindari bias. Kappa dihitung dalam dua tahap: (1) deteksi keberadaan aspek, dan (2) sentimen khusus pada data yang aspeknya disepakati kedua pihak. Nilai Kappa diinterpretasikan menggunakan skala Landis & Koch [32] pada Tabel II.

TABEL II
INTERPRETASI NILAI KAPPA

Nilai Kappa	Tingkat Kesepakatan
< 0,00	Poor
0,00 – 0,20	Slight
0,21 – 0,40	Fair
0,41 – 0,60	Moderate
0,61 – 0,80	Substantial
0,81 – 1,00	Almost Perfect

Selain dihitung menggunakan Cohen's Kappa, performa deteksi aspek (ACD) pada sampel validasi ini juga dievaluasi secara formal menggunakan metrik *Precision*, *Recall*, dan *F1-Score*, dengan memperlakukan label hasil anotasi manual sebagai *gold-standard* dan hasil deteksi aspek oleh Gemini sebagai prediksi. Evaluasi ini dilakukan secara terpisah untuk setiap aspek maupun secara keseluruhan (*macro-average* dan *micro-average*), guna melengkapi ukuran kesepakatan Kappa dengan metrik performa deteksi yang lebih umum digunakan dalam evaluasi tugas klasifikasi biner.

D. Stratified K-Fold Cross Validation

Pembagian data pada masing-masing aspek dilakukan menggunakan teknik *Stratified K-Fold Cross Validation* dengan jumlah *fold* sebanyak $K = 5$, yang digunakan sebagai skema pembagian data *train* dan *test* pada tahap fine-tuning IndoBERT maupun evaluasi model *baseline* (SVM dan NB). Berbeda dengan K-Fold biasa yang membagi data secara acak murni, *Stratified K-Fold* memastikan bahwa proporsi setiap kelas label (Negatif/Netral/Positif) tetap seimbang pada setiap *fold*, baik pada bagian *train* maupun *test*, sehingga mencegah bias evaluasi akibat distribusi kelas yang tidak seimbang antaraspek dalam penelitian ini. Kohavi menunjukkan melalui eksperimen skala besar bahwa *stratified K-fold cross validation* menghasilkan estimasi akurasi yang lebih stabil dibandingkan *cross validation* biasa [33]. Pemilihan $K=5$ pada penelitian ini mengikuti rekomendasi James et al. yang menyatakan bahwa $K=5$ atau $K=10$ memberikan keseimbangan *bias-variance* yang baik secara empiris, khususnya untuk ukuran dataset yang tidak terlalu besar [34].

E. Hyperparameter Tuning

Tahap *hyperparameter tuning* dilakukan menggunakan *framework Optuna*, sebuah kerangka kerja optimasi berbasis pendekatan Bayesian yang mencari kombinasi parameter

optimal secara lebih efisien dibandingkan *grid search* maupun *random search*, dengan memfokuskan pencarian pada area ruang parameter yang paling menjanjikan berdasarkan hasil trial sebelumnya [35]. Tiga parameter utama yang dioptimasi adalah *learning rate*, *dropout*, dan *batch size*, yang dicari secara terpisah untuk setiap aspek berdasarkan nilai F1-score terbaik dari 10 kali percobaan (trial), sehingga diperoleh konfigurasi *hyperparameter* yang paling sesuai dengan karakteristik data pada masing-masing aspek sebelum digunakan pada tahap *fine-tuning* penuh.

F. Fine-Tuning IndoBERT

Tahap fine-tuning IndoBERT final dilakukan pada masing-masing aspek menggunakan skema *5-fold cross-validation*. Proses *fine-tuning* dilakukan secara *end-to-end* dengan menambahkan lapisan klasifikasi (*classification head*) di atas *backbone* IndoBERT, menggunakan konfigurasi *hyperparameter* terbaik hasil optimasi Optuna pada tahap sebelumnya, serta mekanisme *early stopping* (*patience* = 2) untuk mencegah *overfitting*. Devlin dkk. menjelaskan bahwa penambahan satu lapisan klasifikasi di atas representasi token [CLS] pada arsitektur BERT memungkinkan proses fine-tuning *end-to-end* tanpa memerlukan rekayasa fitur tambahan [36]. Performa model pada setiap *fold* dievaluasi menggunakan metrik *Accuracy*, *Precision*, *Recall*, dan F1-Score, serta divisualisasikan melalui kurva *loss* dan *confusion matrix* untuk mengamati konvergensi model dan pola kesalahan klasifikasi. Model dengan F1-score tertinggi pada setiap aspek disimpan sebagai model terbaik, sedangkan hasil evaluasi dari kelima *fold* dirata-ratakan untuk memperoleh performa akhir model pada setiap aspek.

Selain evaluasi *cross-validation* pada data yang dilabeli LLM Gemini, model terbaik pada setiap aspek juga dievaluasi tambahan pada subset *gold-standard*, yaitu bagian dari 1.500 sampel validasi yang telah memiliki label sentimen hasil anotasi manual sepenuhnya oleh manusia. Evaluasi ini bertujuan memastikan performa tinggi model tidak semata disebabkan oleh kesamaan pola antara prediksi model dan label LLM Gemini yang digunakan selama pelatihan, menggunakan metrik *Accuracy*, *Precision*, *Recall*, dan F1-Score (*macro-average*).

G. Perbandingan Model

Tahap ini dilakukan untuk membandingkan performa model IndoBERT dengan dua kelompok model pembanding, yaitu model *baseline* konvensional dan model *transformer* alternatif. Kelompok pertama terdiri atas SVM (*kernel linear*) dan *Naive Bayes* (*Multinomial*) berbasis fitur TF-IDF, guna mengukur peningkatan performa model *transformer* dibandingkan pendekatan *machine learning* tradisional dalam klasifikasi sentimen berbasis aspek. Kedua model *baseline* tersebut diimplementasikan menggunakan modul SVM dan *Naive Bayes* pada *library scikit-learn* dengan parameter *default* ($C=1.0$ untuk SVM; $\alpha=1.0$, $\text{fit_prior}=\text{True}$ untuk *Naive Bayes*) [37]. Kelompok kedua adalah IndoRoBERTa (*flax-community/indonesian-roberta-base*), sebuah model

transformer berbasis arsitektur RoBERTa yang dilatih pada korpus bahasa Indonesia, digunakan sebagai pembanding untuk mengukur apakah performa IndoBERT konsisten unggul dibandingkan varian arsitektur transformer lain pada domain yang sama. IndoRoBERTa dilatih menggunakan skema 5-fold *cross-validation* yang identik dengan IndoBERT, yaitu pembagian *train/test* yang sama persis pada setiap fold per aspek, agar perbandingan bersifat adil (*apple-to-apple*). Model ini menggunakan konfigurasi hyperparameter default yang wajar (*learning rate*= 2×10^{-5} , *dropout*=0,2, *batch size*=16, *epoch*=5).

Keempat model dievaluasi menggunakan skema 5-fold *cross-validation* yang identik untuk setiap aspek, dengan metrik *Accuracy*, *Precision*, *Recall*, dan F1-Score, di mana F1-Score dijadikan metrik utama karena lebih representatif dalam mengukur keseimbangan *precision* dan *recall* pada distribusi kelas yang tidak seimbang. Hasil perbandingan divisualisasikan melalui grafik F1-Score per aspek untuk masing-masing model, sehingga dapat teridentifikasi model yang paling unggul pada setiap aspek.

H. Knowledge Discovery

Peneliti melaksanakan tahap *Knowledge Discovery* setelah memperoleh model klasifikasi terbaik pada tahap sebelumnya untuk menggali wawasan yang bermakna dari hasil prediksi aspek dan sentimen pada ulasan pengguna aplikasi ShopeePay [12]. Pendekatan serupa telah diterapkan pada data tekstual ulasan aplikasi [38] maupun data numerik volume sampah kota [39]. Tahap ini mencakup dua langkah:

1) *Eksplorasi tema*: eksplorasi tema menggunakan ekstraksi kata kunci berbasis TF-IDF dan *topic modeling* dengan *Latent Dirichlet Allocation* (LDA) untuk menggali pola permasalahan spesifik di balik sentimen [13], sebagaimana juga dilakukan melalui analisis Word Cloud pada ulasan aplikasi perbankan digital [40].

2) *Pengambilan sampel ulasan representative*: sebagai bukti kualitatif yang memperjelas hasil analisis kuantitatif dan tematik pada dua langkah sebelumnya.

III. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Proses pengumpulan data yang dilakukan melalui teknik *web scraping* menggunakan pustaka *google-play-scraper* berhasil mengumpulkan sebanyak 42.145 ulasan pengguna aplikasi ShopeePay dari Google Play Store yang diunggah dalam rentang waktu Januari hingga Juni 2026. Data yang berhasil dikumpulkan terdiri atas tiga kolom utama, yaitu rating (skor penilaian pengguna berskala 1–5), review (teks ulasan yang ditulis pengguna), dan tanggal (waktu ulasan tersebut dipublikasikan).

B. Pra-pemrosesan Data

Tahap data pra-pemrosesan bertujuan untuk membersihkan dan mempersiapkan data agar dapat diproses oleh model pembelajaran mesin. Proses ini meliputi *cleaning*, *case*

folding, *tokenization*. Pada penelitian ini, proses *stemming* dan *stopword removal* tidak dilakukan karena model berbasis transformer seperti IndoBERT memanfaatkan informasi kontekstual dari setiap kata, sehingga penghapusan kata atau perubahan bentuk kata berpotensi mengurangi kualitas representasi semantik yang dipelajari model. Tabel III menyajikan contoh hasil setiap tahap prapemrosesan.

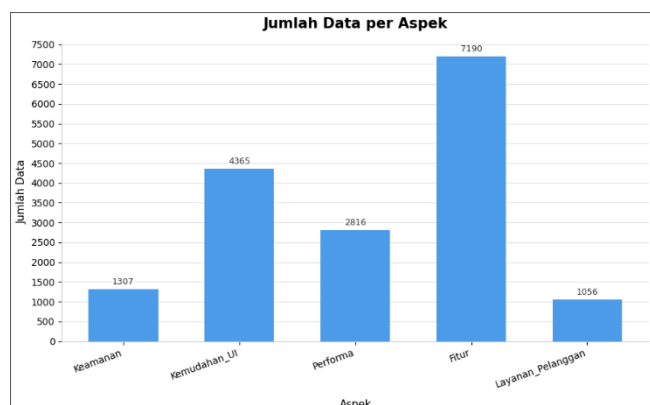
TABEL III
HASIL PRAPEMROSESAN DATA

No	Proses	Hasil
1	Teks Awal	pengen log in ga bisa alasan keamanan padahal udah jelas, ribett banget
2	<i>Cleaning</i>	pengen log in ga bisa alasan keamanan padahal udah jelas ribett banget
3	<i>Case folding</i>	pengen log in ga bisa alasan keamanan padahal udah jelas ribett banget
4	Normalisasi	ingin log in tidak bisa alasan keamanan padahal sudah jelas susah sangat
5	<i>Tokenize</i>	['[CLS]', 'ingin', 'log', 'in', 'tidak', 'bisa', 'alasan', 'keamanan', 'padahal', 'sudah', 'jelas', 'susah', 'sangat', '[SEP]', '[PAD]', ..., '[PAD]'] [2, 561, 1776, 48, 119, 166, 1954, 2957, 2234, 259, 1127, 3684, 310, 3, 0, ..., 0] [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, ..., 0]

Tabel menunjukkan contoh hasil tahapan prapemrosesan. Melalui rangkaian tahapan tersebut, teks ulasan yang awalnya tidak baku dan mengandung *noise* berhasil diseragamkan formatnya menjadi bentuk yang lebih bersih, konsisten, dan siap diproses secara numerik. Setelah keseluruhan tahapan prapemrosesan diterapkan pada 42.145 data mentah hasil *scraping*, diperoleh jumlah data final sebanyak 22.756 ulasan unik, yang selanjutnya digunakan sebagai dasar untuk tahap pelabelan aspek.

C. Pelabelan Data

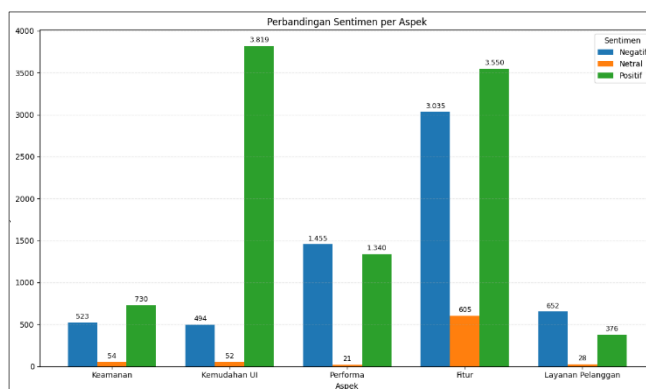
Pada tahap ini, setiap ulasan dari 22.756 data hasil prapemrosesan dianalisis untuk mendeteksi keberadaan lima aspek yang menjadi fokus penelitian, yaitu Keamanan, Kemudahan UI, Performa, Fitur, dan Layanan Pelanggan. Dari 22.756 data, sebanyak 12.525 ulasan (55,0%) teridentifikasi mengandung minimal satu aspek, sedangkan 10.231 ulasan (45,0%) tidak membahas aspek apa pun. Ulasan tanpa aspek tetap dipertahankan dalam dataset hasil pelabelan karena merepresentasikan kondisi nyata pengguna yang memberi opini umum tanpa menyebut aspek layanan tertentu (misalnya "bagus", "mantap"), penghapusannya berpotensi menimbulkan bias seleksi dan menghilangkan perannya sebagai instance negatif pada evaluasi klasifikasi aspek (ACD). Meskipun demikian, ulasan tanpa aspek tidak diikutsertakan pada tahap klasifikasi sentimen (ASC) karena tidak memiliki label sentimen pada aspek manapun.



Gambar 2. Distribusi Label Aspek

Distribusi jumlah data untuk masing-masing aspek disajikan pada Gambar 2 menunjukkan bahwa aspek Fitur memperoleh jumlah data terbanyak, yaitu 7.190 data, diikuti oleh aspek Kemudahan UI sebanyak 4.365 data dan aspek Performa sebanyak 2.816 data. Aspek Keamanan dan Layanan Pelanggan memperoleh jumlah data yang jauh lebih sedikit, yaitu 1.307 dan 1.056 data. Pola ini mengindikasikan bahwa pengguna aplikasi cenderung lebih sering menyampaikan opini terkait fitur transaksi dan kemudahan UI aplikasi dibandingkan aspek keamanan maupun layanan pelanggan. Perbedaan jumlah data yang cukup signifikan antar-aspek ini menandakan adanya ketidakseimbangan data, sehingga peneliti perlu memperhatikan hal ini pada tahap pemodelan selanjutnya, khususnya dalam penggunaan skema *cross-validation* dan pemilihan metrik evaluasi.

Pada tahap pelabelan sentimen, setiap aspek dilabeli tiga kategori sentimen, yaitu positif, negatif, dan netral.



Gambar 3. Perbandingan Sentimen Per Aspek

Aspek Kemudahan UI dan Fitur memperoleh dominasi sentimen positif pada Gambar 3, dengan 3.819 dan 3.550 ulasan, sehingga mayoritas pengguna menunjukkan kepuasan terhadap kemudahan penggunaan antarmuka aplikasi. Aspek Performa memperoleh dominasi sentimen negatif, dengan 1.455 ulasan, sehingga aspek ini mengindikasikan keluhan pengguna terkait kestabilan performa aplikasi. Aspek Keamanan cenderung lebih banyak disertai sentimen positif

(730 ulasan) dibandingkan negatif (523 ulasan), sedangkan aspek Layanan Pelanggan menunjukkan kecenderungan sebaliknya, dengan sentimen negatif (652 ulasan) yang lebih banyak dibandingkan positif (376 ulasan) mengindikasikan bahwa pengguna relatif puas terhadap sisi keamanan aplikasi, namun cenderung mengeluhkan penanganan dari layanan pelanggan.

Untuk memastikan kualitas dan konsistensi hasil pelabelan yang dilakukan, dilaksanakan proses validasi label dengan membandingkan hasil pelabelan antara penulis dan Gemini, menggunakan Cohen's Kappa dan persentase agreement pada 1.500 data sampel hasil *stratified sampling*. Hasil validasi disajikan pada Tabel IV.

TABEL IV
HASIL VALIDASI LABEL

	Cohen's Kappa	% Agreement	Interpretasi
Aspek	0,744	92,6%	Substantial
Sentimen	0,903	95%	Almost Perfect

Penulis dan Gemini mencapai tingkat kesepakatan kategori *Substantial* dalam mendeteksi keberadaan aspek, dengan rata-rata Kappa sebesar 0,744 dan %Agreement sebesar 92,6%. Penulis dan Gemini juga mencapai tingkat kesepakatan kategori *Almost Perfect* dalam pelabelan sentimen, dengan rata-rata Kappa sebesar 0,903 dan %Agreement sebesar 95%. Berbeda dengan asumsi umum bahwa klasifikasi sentimen cenderung lebih subjektif dibandingkan deteksi aspek, pada penelitian ini justru diperoleh pola sebaliknya, kesepakatan sentimen lebih tinggi dibandingkan kesepakatan deteksi aspek. Hal ini kemungkinan disebabkan oleh perhitungan Kappa sentimen yang hanya dilakukan pada subset data yang sudah disepakati kedua pihak memiliki aspek tersebut (*agreed has-aspect*), sehingga ulasan dengan sentimen ambigu di batas keputusan "ada aspek atau tidak" sudah tersaring lebih dulu pada tahap deteksi aspek, menyisakan kasus sentimen yang relatif lebih jelas polaritasnya untuk dinilai.

Pada level aspek individual, ditemukan variasi kesepakatan yang cukup signifikan antar aspek. Aspek Keamanan (Kappa=0,862), Layanan Pelanggan (Kappa=0,811), Kemudahan UI (Kappa=0,782) dan Performa (Kappa=0,714) berada pada kategori *Almost Perfect* hingga *Substantial* tinggi. Sebaliknya, aspek Fitur memperoleh Kappa terendah (0,555, kategori *Moderate*) meskipun %Agreement-nya masih tergolong tinggi (80,3%). Perbedaan ini terjadi karena Kappa mengoreksi kemungkinan kesepakatan akibat kebetulan (*chance agreement*), dan aspek Fitur memiliki proporsi kelas positif (ulasan yang membahas Fitur) yang jauh lebih besar dibandingkan aspek lain, sehingga koreksi *chance agreement*-nya lebih besar pula. Hal ini sejalan dengan sifat aspek Fitur yang mencakup cakupan topik paling luas di antara kelima aspek (mencakup berbagai fitur transaksi seperti transfer, top up, QRIS, dan ShopeePayLater), sehingga berpotensi menimbulkan lebih banyak perbedaan interpretasi batas topik antara Gemini dan anotator manual dibandingkan aspek lain yang cakupannya lebih sempit dan spesifik.

Selain kesepakatan Kappa, performa deteksi aspek (ACD) juga dievaluasi menggunakan metrik *Precision*, *Recall*, dan *F1-Score* terhadap *gold-standard* anotasi manual, sebagaimana disajikan pada Tabel V.

TABEL V
HASIL EVALUASI ASPECT CATEGORY DETECTION (ACD)

Aspek	Precision	Recall	F1
Keamanan	0,909	0,833	0,870
Kemudahan UI	0,864	0,788	0,825
Performa	0,868	0,668	0,755
Fitur	0,707	0,697	0,702
Layanan Pelanggan	0,877	0,770	0,820
Macro-average	0,845	0,751	0,794
Micro-average	0,799	0,728	0,762

Berdasarkan Tabel V, model deteksi aspek berbasis Gemini mencapai F1-Score *macro-average* sebesar 0,794 terhadap *gold-standard* anotasi manual. Aspek Keamanan memperoleh performa ACD tertinggi (F1=0,870), sedangkan aspek Fitur memperoleh performa terendah (F1=0,702), konsisten dengan temuan nilai Kappa terendah pada aspek yang sama (Kappa=0,555), yang mengonfirmasi bahwa cakupan topik yang luas pada aspek Fitur menyulitkan penarikan batas keputusan "ada aspek atau tidak" secara konsisten. Aspek Performa menunjukkan pola *precision* tinggi namun *recall* relatif rendah (0,868 berbanding 0,668), mengindikasikan bahwa model deteksi cenderung konservatif dan melewatkan sejumlah ulasan yang sebenarnya membahas aspek Performa (*false negative*) dibandingkan salah mendeteksi aspek yang tidak relevan. Nilai F1-Score ACD ini menunjukkan bahwa tahap deteksi aspek (ACD) merupakan sumber kesalahan yang lebih dominan dalam keseluruhan pipeline ACSA dibandingkan tahap klasifikasi sentimennya (ASC). Tabel VI menampilkan sampel data hasil pelabelan.

Proses pelabelan sentimen menjalani tahap penghapusan lanjutan setelah validasi Kappa selesai dilakukan, yaitu menghapus baris data dengan label sentimen "Netral" pada masing-masing aspek, karena penelitian ini berfokus pada klasifikasi sentimen biner (Positif dan Negatif).

TABEL VI
SAMPEL DATA HASIL PELABELAN

Content	Keamanan	Kemudahan UI	Performa	Fitur	Layanan Pelanggan
aplikasi sering ngelag dan blank hmmm	None	None	Negatif	None	None
fitur spinjam tidak bisa diakses selalu terjadi kesalahan halaman ketika di buka	None	None	Negatif	Negatif	None

Selain menjaga proporsi kelas sentimen dalam setiap fold, ketidakseimbangan jumlah total data antar-aspek ini diatasi dengan melakukan proses *fine-tuning* dan optimasi *hyperparameter* secara terpisah untuk setiap aspek, sehingga performa model pada aspek dengan data lebih sedikit tetap dievaluasi menggunakan konfigurasi yang paling sesuai dengan karakteristik datanya masing-masing, bukan disamaratakan dengan aspek bervolume data besar.

D. Stratified K-Fold Cross Validation

Dataset final dibagi menggunakan metode *stratified k-fold cross-validation* dengan nilai $k=5$ sebelum proses pelatihan model dilakukan. Teknik ini mempertahankan proporsi kelas Positif dan Negatif pada setiap fold agar tetap merepresentasikan distribusi kelas pada dataset asli, sehingga teknik ini mengurangi risiko bias akibat pembagian data yang tidak merata. Peneliti menerapkan pembagian *fold* pada tingkat ulasan (dokumen), bukan pada tingkat klausa. Skema ini mencegah *data leakage* yang dapat terjadi apabila klausa dari satu ulasan tersebar pada *fold* yang berbeda. Tabel VII menyajikan jumlah total data serta rata-rata jumlah data latih dan data uji pada setiap fold untuk kelima aspek yang diuji.

TABEL VII
DISTRIBUSI DATA STRATIFIED 5-FOLD PER ASPEK

Aspek	Total Data (Biner)	Rata-rata Data Latih/Fold	Rata-rata Data Uji/Fold
Keamanan	1.253	1.002	251
Kemudahan UI	4.313	3.450	863
Performa	2.795	2.236	559
Fitur	6.597	5.277	1.319
Layanan Pelanggan	1.028	822	206

Kelima aspek menghasilkan jumlah data biner yang berbeda-beda karena setiap aspek memiliki jumlah anotasi yang tidak sama. Skema *cross-validation* membagi data dengan proporsi data latih dan data uji sekitar 80:20 pada setiap fold untuk masing-masing aspek. Aspek Fitur menghasilkan jumlah data biner terbanyak, yaitu 6.597 data, diikuti oleh aspek Kemudahan UI dengan 4.313 data. Aspek Layanan Pelanggan menghasilkan jumlah data biner paling sedikit, yaitu hanya 1.028 data, sehingga proses pelatihan model perlu memperhatikan potensi ketidakseimbangan data (*data imbalance*) antar-aspek, khususnya antara aspek Layanan Pelanggan dan aspek Fitur yang selisih jumlah datanya mencapai lebih dari 5.000 data.

E. Hyperparameter Tuning

Proses *hyperparameter tuning* dilakukan untuk mencari kombinasi parameter terbaik pada model IndoBERT dalam melakukan klasifikasi sentimen untuk masing-masing aspek. Parameter yang dioptimasi meliputi *learning rate*, *dropout*, dan *batch size*, karena ketiga parameter ini sangat berpengaruh terhadap kemampuan model dalam belajar pola

dari data tanpa mengalami overfitting maupun underfitting. Tabel VIII menyajikan hasil kombinasi *hyperparameter* terbaik yang diperoleh untuk masing-masing aspek berdasarkan proses optimasi menggunakan Optuna.

TABEL VIII
HASIL HYPERPARAMETER TERBAIK PER ASPEK

Aspek	Learning Rate	Dropout	Batch Size
Keamanan	$2,10 \times 10^{-5}$	0,219	8
Kemudahan UI	$2,59 \times 10^{-5}$	0,222	8
Performa	$1,25 \times 10^{-5}$	0,189	16
Fitur	$1,76 \times 10^{-5}$	0,340	16
Layanan Pelanggan	$2,44 \times 10^{-5}$	0,182	8

Learning rate pada kelima aspek berkisar antara $1,25 \times 10^{-5}$ hingga $2,59 \times 10^{-5}$, sementara *dropout* berkisar antara 0,182 hingga 0,340. Berbeda dengan asumsi umum bahwa dataset berukuran kecil membutuhkan nilai *dropout* yang lebih besar untuk mencegah *overfitting*, hasil optimasi pada penelitian ini justru menunjukkan pola sebaliknya: aspek Fitur, yang memiliki jumlah data biner terbanyak (6.597 data), memperoleh nilai *dropout* tertinggi (0,340), sedangkan aspek Layanan Pelanggan, yang memiliki jumlah data biner paling sedikit (1.028 data), justru memperoleh nilai *dropout* terendah (0,182). Aspek Keamanan, yang juga memiliki jumlah data relatif kecil (1.253 data), memperoleh nilai *dropout* menengah (0,219), tidak jauh berbeda dari aspek Kemudahan UI (0,222) yang jumlah datanya lebih dari tiga kali lipat lebih besar. Temuan ini menunjukkan bahwa kombinasi *hyperparameter* optimal tidak berkorelasi secara sederhana dan monoton dengan ukuran data pada masing-masing aspek, melainkan lebih dipengaruhi oleh interaksi kompleks antar-parameter serta karakteristik distribusi data (misalnya keseimbangan kelas Positif/Negatif dan variasi linguistik ulasan) yang ditemukan oleh proses pencarian *Bayesian Optuna* berdasarkan performa pada *fold* validasi. Hal ini menegaskan pentingnya proses optimasi *hyperparameter* secara individual pada setiap aspek, alih-alih menggunakan asumsi umum atau nilai *default* yang seragam, untuk mencapai performa klasifikasi sentimen yang optimal pada model IndoBERT.

F. Fine-Tuning IndoBERT

Setelah nilai *hyperparameter* optimal diperoleh, penelitian ini melakukan proses *fine-tuning* model IndoBERT (*indobert-base-p2*) pada skema 5-fold *cross-validation* untuk setiap aspek, dengan *epoch* maksimum sebanyak 15 dan pengaturan *early stopping* untuk mencegah *overfitting*. Tabel IX menyajikan hasil F1-score pada setiap fold beserta rata-ratanya.

TABEL IX
HASIL F1-SCORE MODEL INDOBERT PADA SETIAP FOLD

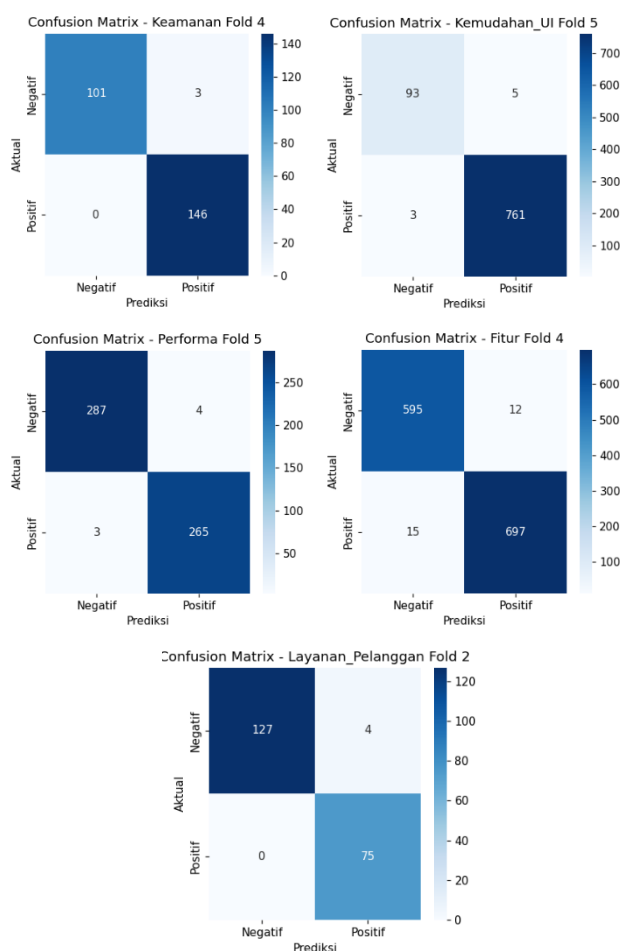
Aspek	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Avg
Keamanan	0,984	0,976	0,984	0,988	0,980	0,982
Kemudahan UI	0,975	0,964	0,971	0,965	0,977	0,970
Performa	0,986	0,987	0,977	0,980	0,987	0,984
Fitur	0,979	0,968	0,979	0,979	0,979	0,977
Layanan Pelanggan	0,958	0,979	0,974	0,979	0,979	0,974

Model mencatatkan rentang F1 terbesar pada aspek Layanan Pelanggan, yaitu dari 0,958 hingga 0,979, dengan standar deviasi sebesar 0,009. Model juga menunjukkan variasi yang tinggi pada aspek Kemudahan UI, dengan rentang F1 dari 0,964 hingga 0,977 dan standar deviasi 0,006. Sebaliknya, model menghasilkan variasi terkecil dan setara pada aspek Keamanan, Performa, dan Fitur, dengan standar deviasi masing-masing sebesar 0,005. Kecilnya variasi antar-*fold* pada ketiga aspek tersebut mengindikasikan bahwa model melakukan generalisasi secara stabil, sedangkan variasi yang lebih besar pada aspek Layanan Pelanggan kemungkinan disebabkan oleh keterbatasan jumlah data pada aspek tersebut, sehingga performa model pada *fold* tertentu lebih rentan terhadap variasi komposisi data uji. Selanjutnya, Tabel X menyajikan rata-rata hasil evaluasi lengkap (*accuracy*, *precision*, *recall*, dan F1-score) dari 5-fold *cross-validation* untuk setiap aspek.

TABEL X
RATA-RATA HASIL 5-FOLD CROSS-VALIDATION MODEL INDOBERT PER ASPEK

Aspek	Accuracy	Precision	Recall	F1-Score
Keamanan	0,982	0,981	0,983	0,982
Kemudahan UI	0,988	0,972	0,969	0,970
Performa	0,984	0,984	0,983	0,984
Fitur	0,977	0,976	0,977	0,977
Layanan Pelanggan	0,976	0,974	0,974	0,974

Berdasarkan Tabel X, model IndoBERT mencapai nilai *Accuracy* dan F1-Score di atas 97% pada seluruh aspek. Model memperoleh *Accuracy* tertinggi pada aspek Kemudahan UI, yaitu sebesar 0,988. Model memperoleh F1-Score tertinggi pada aspek Performa, yaitu sebesar 0,984, sehingga aspek tersebut menunjukkan keseimbangan *precision* dan *recall* yang paling baik. Model memperoleh F1-Score terendah pada aspek Kemudahan UI, yaitu sebesar 0,970, menariknya, aspek ini justru memiliki *Accuracy* tertinggi namun F1-Score terendah, yang mengindikasikan adanya sedikit ketidakseimbangan antara *precision* (0,972) dan *recall* (0,969) pada aspek tersebut, meski selisihnya kecil. Selisih F1-Score antar-aspek berkisar kurang dari 1,5%, sehingga model IndoBERT menunjukkan kemampuan generalisasi yang baik pada seluruh aspek meskipun jumlah data pada setiap aspek berbeda-beda secara signifikan.



Gambar 4. Confusion Matrix

Penelitian ini juga melakukan analisis terhadap *confusion matrix* dari model IndoBERT pada *fold* terbaik masing-masing aspek, sebagai representasi performa model dalam tugas klasifikasi sentimen per aspek. Hasil confusion matrix disajikan pada Gambar 4 untuk kelima aspek, yaitu Keamanan, Kemudahan UI, Performa, Fitur, dan Layanan Pelanggan. Model memperoleh jumlah prediksi benar yang sangat tinggi secara konsisten pada aspek Keamanan, Kemudahan UI, dan Performa, dengan tingkat kesalahan masing-masing di bawah 1,3% dari total data uji. Model menunjukkan tingkat kesalahan tertinggi pada aspek Fitur, dengan total 27 kesalahan prediksi dari 1.319 data uji (2,0%), terdiri atas 12 ulasan Negatif yang salah diprediksi sebagai Positif (*false positive*) dan 15 ulasan Positif yang salah diprediksi sebagai Negatif (*false negative*). Berbeda dengan aspek lain yang kesalahannya cenderung berat sebelah ke satu arah, aspek Fitur menunjukkan jumlah *false positive* dan *false negative* yang relatif seimbang, mengindikasikan bahwa model kesulitan membedakan sentimen pada kedua arah secara merata pada aspek ini, konsisten dengan sifat aspek Fitur yang mencakup topik paling luas di antara kelima aspek. Aspek Layanan Pelanggan menunjukkan tingkat kesalahan

tertinggi kedua (1,9%), namun seluruh kesalahannya (4 dari 206 data) berupa *false positive* (ulasan Negatif diprediksi Positif), tanpa satu pun *false negative*. Kecenderungan kesalahan yang mengarah ke *false positive* pada aspek ini mengindikasikan bahwa model belum sepenuhnya maksimal dalam mengenali sentimen negatif pada topik layanan pelanggan, kemungkinan karena keluhan terhadap CS sering disampaikan dengan nada yang lebih halus atau tersirat dibandingkan keluhan pada aspek teknis seperti Performa.

Sebagai validasi tambahan, model IndoBERT terbaik pada setiap aspek juga diuji langsung terhadap subset data gold-standard (anotasi manual), sebagai pembandingan tambahan di luar label LLM Gemini yang digunakan selama pelatihan. Tabel XI menyajikan hasil evaluasi tersebut beserta perbandingannya terhadap F1-Score cross-validation.

TABEL XI
PERBANDINGAN F1-SCORE: CROSS-VALIDATION (LABEL LLM GEMINI) vs GOLD-STANDARD (MANUAL)

Aspek	N	F1 CV (Label LLM Gemini)	F1 Gold-Standard	Selisih
Keamanan	72	0,982	0,971	+0,011
Kemudahan UI	306	0,970	0,974	-0,004
Performa	247	0,984	0,991	-0,007
Fitur	462	0,977	0,980	-0,003
Layanan Pelanggan	67	0,974	1,000	- 0,026

Tabel XI menunjukkan hasil evaluasi pada lima aspek. Aspek Kemudahan UI, Performa, Fitur, dan Layanan Pelanggan memperoleh F1-Score lebih tinggi pada data gold-standard dibandingkan pada data cross-validation berlabel LLM Gemini, dengan selisih berkisar antara 0,003 hingga 0,026. Hanya aspek Keamanan yang menunjukkan pola sebaliknya, dengan selisih yang relatif kecil (0,011). Pola ini mengindikasikan bahwa performa tinggi model tidak disebabkan oleh model semata "menghafal" pola dari label LLM Gemini yang digunakan selama pelatihan, melainkan turut divalidasi terhadap penilaian sentimen oleh manusia. Perlu dicatat bahwa selisih terbesar terjadi pada aspek Layanan Pelanggan, yang memiliki ukuran sampel gold-standard paling kecil (N=67) dan mencapai F1 sempurna (1,000); hasil ini sebaiknya diinterpretasikan secara hati-hati mengingat potensi varians tinggi akibat ukuran sampel terbatas, bukan sebagai bukti generalisasi sempurna pada populasi data yang lebih luas.

G. Perbandingan Model

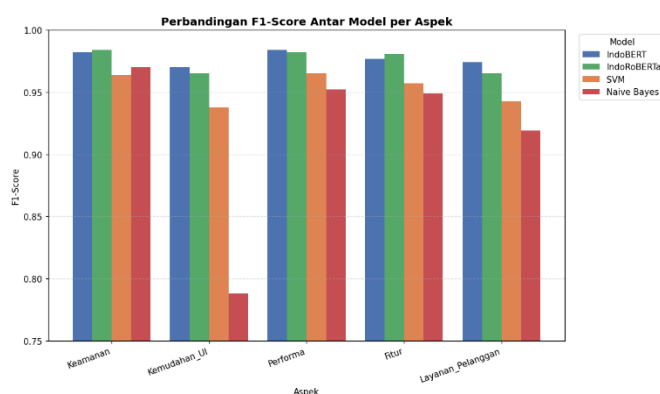
Penelitian ini membandingkan performa (F1-score) antara model IndoBERT, IndoRoBERTa, SVM, dan *Naive Bayes* pada masing-masing aspek. Tabel XII menyajikan hasil perbandingan performa keempat model tersebut secara lengkap untuk setiap aspek yang diuji.

TABEL XII
HASIL PERBANDINGAN MODEL

Aspek	IndoBERT	IndoRoBERTa	SVM	Naive Bayes
Keamanan	0,982	0,984	0,964	0,970
Kemudahan UI	0,970	0,965	0,938	0,788
Performa	0,984	0,982	0,965	0,952
Fitur	0,977	0,981	0,957	0,949
Layanan Pelanggan	0,974	0,965	0,943	0,919

Kedua model berbasis *transformer* (IndoBERT dan IndoRoBERTa) secara konsisten mengungguli kedua model *baseline* konvensional (SVM dan *Naive Bayes*) pada seluruh aspek yang diuji, dengan selisih F1-score yang cukup besar terutama terhadap *Naive Bayes*. Namun demikian, di antara kedua model *transformer* tersebut, tidak ada satu model yang unggul mutlak pada seluruh aspek. IndoRoBERTa memperoleh F1-score sedikit lebih tinggi dibandingkan IndoBERT pada aspek Keamanan (0,984 berbanding 0,982) dan Fitur (0,981 berbanding 0,977), sedangkan IndoBERT tetap unggul pada aspek Kemudahan UI, Performa, dan Layanan Pelanggan. Perlu dicatat bahwa IndoRoBERTa pada penelitian ini menggunakan konfigurasi *hyperparameter default* tanpa melalui proses *tuning* terpisah menggunakan Optuna sebagaimana dilakukan pada IndoBERT.

Naive Bayes mencatatkan selisih performa paling besar terhadap model *transformer* pada aspek Kemudahan UI, dengan F1-score hanya 0,788 jauh di bawah performanya pada aspek lain (0,919–0,970). Asumsi independensi antar-fitur pada *Naive Bayes* terlalu sederhana untuk menangkap kompleksitas bahasa pada ulasan yang membahas kemudahan penggunaan. Hasil ini secara keseluruhan menunjukkan bahwa pendekatan *deep learning* berbasis *transformer* menangkap konteks dan makna semantik dalam teks ulasan berbahasa Indonesia secara lebih baik dibandingkan model *machine learning* konvensional yang mengandalkan representasi fitur seperti TF-IDF. Keunggulan di antara kedua varian *transformer* itu sendiri tetap bergantung pada karakteristik masing-masing aspek.



Gambar 5. Hasil Perbandingan Model

Hasil pada Gambar 5 membuktikan secara keseluruhan bahwa pendekatan berbasis *pre-trained language model* (IndoBERT dan IndoRoBERTa) memberikan peningkatan performa yang signifikan dan konsisten dibandingkan model *machine learning* tradisional dalam tugas klasifikasi sentimen berbasis aspek, khususnya pada data dengan kompleksitas bahasa yang lebih tinggi seperti aspek Kemudahan UI. Hal ini menegaskan pentingnya representasi kontekstual yang dimiliki oleh model *transformer* dalam memahami nuansa bahasa Indonesia pada teks ulasan pengguna.

H. Knowledge Discovery

Berdasarkan hasil klasifikasi sentimen pada kelima aspek yang diteliti, mengungkap bahwa aspek Fitur dan Performa merupakan dua aspek dengan sentimen negatif tertinggi sebanyak 3.035 dan 1.455 ulasan. Temuan ini mengindikasikan keluhan pengguna terkait kelengkapan fitur serta kestabilan performa aplikasi seperti kendala transaksi, loading lambat, atau error sistem. Perlu dicatat bahwa tingginya jumlah keluhan pada aspek Fitur tidak serta-merta berarti sentimen aspek ini secara keseluruhan negatif. Secara proporsi, aspek Fitur tetap didominasi sentimen positif, namun karena volume ulasannya jauh lebih besar dibanding aspek lain, jumlah ulasan negatifnya pun ikut menjadi yang tertinggi secara nominal. Sebaliknya, aspek Kemudahan UI menunjukkan proporsi sentimen positif tertinggi, sementara aspek Keamanan dan Layanan Pelanggan menunjukkan distribusi sentimen yang lebih seimbang antara positif dan negatif. Untuk memperjelas pola distribusi tersebut, dilakukan ekstraksi kata kunci paling menonjol dan *topic modeling* (LDA) pada setiap aspek.

Pada aspek Keamanan, kata kunci sentimen negatif pada aspek ini didominasi oleh "akun", "saldo", "pin", "hilang", "login", dan "lupa". Topik 1 dan 2 hasil LDA ("akun, ktp, data" dan "saldo, pin, pay, malah") memperkuat pola ini dan menunjukkan bahwa keluhan pengguna terkonsentrasi pada dua isu utama, yaitu kehilangan atau pengurangan saldo tanpa transaksi yang jelas, serta kendala akses akun seperti lupa PIN dan kegagalan verifikasi KTP. Sebaliknya, Topik 3 pada sentimen positif ("aman, terpercaya, amanah") mencerminkan kepercayaan pengguna terhadap keamanan transaksi secara umum.

Pola serupa juga ditemukan pada aspek Kemudahan UI, dengan kata kunci negatif pada aspek ini meliputi "ribet", "susah", "verifikasi", "wajah", dan "daftar". Didukung oleh Topik 3 ("ribet, bayar, harus, transfer"), temuan ini menunjukkan bahwa keluhan pengguna berpusat pada proses verifikasi dan registrasi yang dianggap berbelit. Menariknya, Topik 2 ("terkadang, lambat, iklan") mengindikasikan bahwa sebagian kecil ulasan pada aspek ini sebenarnya menyinggung isu performa dan gangguan iklan. Kedua isu tersebut bercampur dengan konteks UI karena sama-sama memengaruhi pengalaman pengguna saat berinteraksi dengan aplikasi.

Sementara itu, aspek Performa menunjukkan kata kunci negatif pada aspek ini meliputi "lambat", "error", "gagal", dan

"lemot". Topik 2 ("lama, terus, masuk, gagal, error") dan Topik 3 ("lambat, bug, dipakai, terjadi") mengonfirmasi bahwa keluhan performa didominasi oleh dua pola berbeda, yaitu kegagalan transaksi atau proses akibat sistem, dan kelambatan aplikasi secara umum. Temuan ini sejalan dengan proporsi sentimen negatif tertinggi kedua yang dimiliki aspek ini.

Berbeda dengan aspek-aspek sebelumnya, aspek Fitur menunjukkan pola keluhan yang tercermin melalui Topik 2 ("bayar, dapat, spinjam, akun") dan Topik 3 ("saldo, dana, transfer, top up, lama"). Kedua topik ini mengungkap dua sumber keluhan utama, yaitu penolakan atau pembekuan pengajuan fitur pinjaman seperti Spaylater, serta kegagalan fungsi transfer atau top up saldo yang tidak sesuai harapan pengguna.

Terakhir, pada aspek Layanan Pelanggan, kata kunci negatif meliputi "cs", "nunggu", "buruk", dan "belum". Didukung oleh Topik 2 ("cs, saldo, hari, dana, nunggu, jam"), temuan ini mengindikasikan bahwa keluhan pengguna dominan berkaitan dengan lambatnya respons dan penyelesaian oleh *customer service*.

Untuk memperjelas pola tersebut, disertakan contoh ulasan representatif untuk tiap kombinasi aspek-sentimen yang disajikan pada tabel XIII.

TABEL XIII
CONTOH ULASAN REPRESENTATIF PER ASPEK DAN SENTIMEN

Aspek	Positif	Negatif
Keamanan	"dompet yang amanah aman puas dan best thank you"	"kenapa akun shopeepay saya tiba tiba non aktif terblokir padahal masih banyak saldonya"
Kemudahan UI	"shopeepay ini sangat sangat berguna dan memudahkan dalam bertransaksi dimana pun itu"	"aplikasinya pilih kasih giliran mau upgrade ke shopeepay plus tidak di bolehin"
Performa	"bagus lancar jaya dan gak ribet cara gunain nya"	"akhir akhir suka ngeblank ngelag harus di hapus data dan cache baru benar lagi"
Fitur	"untuk keseluruhannya sudah bagus mulai dari fitur nya yang lengkap..."	"pengajuan saya kenapa di tolak padahal sudah tandatangan kontrak dan tidak bisa di cairkan"
Layanan Pelanggan	"bagus pelayanan bantuan nya juga cepat dan di respon"	"spaylater di bekukan dan belum ada tanggapan sampe sekarang"

Pola-pola pada Tabel XIII konsisten dengan hasil ekstraksi kata kunci dan *topic modeling* pada langkah sebelumnya, memperkuat bahwa setiap aspek memiliki tema keluhan dan apresiasi yang khas, bukan sekadar variasi jumlah sentimen.

Hal ini memiliki relevansi khusus dalam konteks industri e-wallet di Indonesia, mengingat layanan pembayaran digital sangat bergantung pada kepercayaan dan keandalan sistem

sebagai faktor penentu retensi pengguna. Tingginya proporsi keluhan pada aspek Layanan Pelanggan dan Performa perlu menjadi perhatian serius, karena lambannya penanganan komplain serta kegagalan transaksi pada e-wallet berdampak langsung pada kepercayaan dan dana pengguna, sekaligus berpotensi mendorong perpindahan pengguna (*churn*) ke kompetitor seperti GoPay, DANA, atau OVO, sejalan dengan temuan bahwa *payment convenience* dan *service quality* merupakan dimensi paling berpengaruh terhadap kepuasan pengguna e-wallet [2], [4]. Sementara itu, meskipun aspek Fitur merupakan aspek yang paling banyak dibahas pengguna, proporsi sentimen negatifnya relatif moderat, mengindikasikan bahwa ekspansi fitur ShopeePay menjadi *super-app* pembayaran secara umum telah cukup diimbangi stabilitas dan kemudahan penggunaan, meski keluhan spesifik seperti penolakan pengajuan pinjaman tetap berisiko mengurangi daya saing di tengah persaingan e-wallet yang semakin bertumpu pada ekosistem fitur. Sebaliknya, apresiasi tinggi terhadap Kemudahan UI menunjukkan bahwa desain antarmuka ShopeePay telah memenuhi ekspektasi *usability* yang diakui sebagai dimensi kritis penentu adopsi *mobile payment* [2], [25], sehingga sebaiknya dipertahankan sebagai *competitive advantage*. Sementara itu, distribusi sentimen yang relatif seimbang pada aspek Keamanan mengindikasikan perlunya komunikasi yang lebih transparan terkait mekanisme keamanan, mengingat aspek ini merupakan atribut paling kritis dalam kepercayaan pengguna *mobile payment* [25], [26].

Dengan demikian, hasil ini memberikan wawasan berbasis data bagi pengembang ShopeePay untuk memprioritaskan perbaikan pada aspek Fitur dan Performa, dan mempertahankan kualitas Kemudahan UI, sekaligus menjadi masukan strategis bagi daya saing ShopeePay dalam ekosistem pembayaran digital Indonesia yang terus berkembang.

IV. KESIMPULAN

Penelitian ini berhasil menerapkan *Aspect-Category Sentiment Analysis* (ACSA) menggunakan Fine-Tuned IndoBERT untuk mengklasifikasikan sentimen pada lima aspek layanan ShopeePay, yaitu Keamanan, Kemudahan UI, Performa, Fitur, dan Layanan Pelanggan, dari 22.756 ulasan pengguna Google Play Store hasil web scraping dan preprocessing, di mana 12.525 ulasan (55,0%) teridentifikasi mengandung minimal satu aspek layanan, dengan pelabelan aspek dan sentimen yang dilakukan secara langsung menggunakan model bahasa besar (LLM Gemini 3.5 Flash-Lite) yang tervalidasi terhadap 1.500 sampel anotasi manual secara blind menggunakan Cohen's Kappa pada kategori *Substantial* hingga *Almost Perfect* (κ deteksi aspek = 0,744; κ sentimen = 0,903), serta dievaluasi lebih lanjut menggunakan metrik *Precision*, *Recall*, dan *F1-Score* terhadap *gold-standard* manual pada tahap deteksi aspek (ACD), menghasilkan *F1-Score macro-average* sebesar 0,794 yang mengonfirmasi bahwa tahap deteksi aspek merupakan sumber

kesalahan yang lebih dominan dalam pipeline ACSA dibandingkan tahap klasifikasi sentimennya.

Hasil evaluasi menggunakan stratified 5-fold *cross-validation* menunjukkan bahwa model IndoBERT secara konsisten mengungguli baseline SVM dan Naive Bayes pada seluruh aspek dengan F1-Score berkisar antara 0,970 hingga 0,984, dengan selisih performa paling signifikan terlihat pada aspek Kemudahan UI (0,970 dibandingkan 0,788 untuk Naive Bayes), yang membuktikan keunggulan pendekatan berbasis transformer dalam menangkap konteks dan kompleksitas bahasa Indonesia dibandingkan model *machine learning* konvensional. Dibandingkan dengan IndoRoBERTa sebagai model transformer alternatif berbasis arsitektur RoBERTa, IndoBERT menunjukkan performa yang relatif setara tanpa perbedaan yang signifikan mengindikasikan bahwa performa tinggi IndoBERT konsisten di antara varian arsitektur transformer berbahasa Indonesia, sekaligus mendukung ketepatan pemilihan IndoBERT sebagai model utama penelitian ini. Hasil ini turut divalidasi melalui evaluasi tambahan pada subset data *gold-standard* yang dianotasi sepenuhnya oleh manusia, menghasilkan F1-Score antara 0,971 hingga 1,000 yang konsisten dengan bahkan sedikit melampaui hasil *cross-validation*, mengonfirmasi bahwa performa tinggi model tidak semata disebabkan oleh kesamaan pola dengan label LLM Gemini yang digunakan selama pelatihan.

Hasil *Knowledge Discovery* mengungkap bahwa aspek Fitur dan Performa merupakan aspek dengan tingkat keluhan tertinggi, sementara aspek Kemudahan UI memperoleh apresiasi positif tertinggi dari pengguna, sehingga temuan ini dapat dijadikan dasar rekomendasi berbasis data bagi pengembang ShopeePay untuk memprioritaskan perbaikan stabilitas sistem dan kelengkapan fitur, sekaligus mempertahankan kualitas antarmuka sebagai keunggulan kompetitif. Penelitian ini memberikan kontribusi akademis berupa bukti empiris konsistensi performa Fine-Tuned IndoBERT pada domain ulasan aplikasi *e-wallet* yang sebelumnya masih terbatas dikaji, sekaligus kontribusi praktis berupa wawasan strategis bagi pengembang ShopeePay dalam menghadapi persaingan industri *e-wallet* di Indonesia.

Meskipun demikian, penelitian ini memiliki beberapa keterbatasan yang perlu menjadi perhatian. Skema ACSA yang digunakan membatasi deteksi aspek pada lima kategori yang telah ditetapkan sebelumnya, sehingga opini pengguna yang membahas aspek layanan di luar kelima kategori tersebut tidak tertangkap oleh model dan berpotensi hilang dari analisis *knowledge discovery*. Selain itu, data penelitian ini hanya bersumber dari ulasan pengguna di Google Play Store, sehingga belum mencakup ulasan dari platform lain seperti App Store maupun media sosial, dan temuan yang dihasilkan belum tentu sepenuhnya merepresentasikan persepsi pengguna ShopeePay secara keseluruhan. Berdasarkan keterbatasan tersebut, penelitian selanjutnya disarankan untuk melatih ulang model dengan mengeluarkan data validasi dari skema fold pelatihan guna mengeliminasi risiko *data leakage*, mengeksplorasi pendekatan *open-domain*

aspect extraction untuk mengidentifikasi kategori aspek baru yang belum tercakup dalam skema *predefined* saat ini, memperluas cakupan data dengan menambah jumlah anotasi pada aspek minoritas seperti Keamanan dan Layanan Pelanggan agar evaluasi performa model pada aspek-aspek tersebut lebih representatif, serta mengeksplorasi klasifikasi sentimen netral atau skema multi-kelas secara lebih mendalam guna memperkaya analisis opini pengguna secara lebih komprehensif.

DAFTAR PUSTAKA

- [1] H. D. Ibrahim, N. Heryana, and A. Primajaya, "Analisis Sentimen Pengguna E-Wallet Ovo dan Dana Pada Twitter Dengan Metode Naive Bayes Classifier," *J. Mhs. Tek. Inform.*, vol. 8, no. 5, pp. 9858–9864, 2024, doi: <https://doi.org/10.36040/jati.v8i5.10738>.
- [2] A. S. Ajina, J. M. M. Joudeh, N. N. Ali, A. M. Zamil, and T. N. Hashem, "The effect of mobile-wallet service dimensions on customer satisfaction and loyalty: An empirical study," *Cogent Bus. Manag.*, vol. 10, no. 2, 2023, doi: [10.1080/23311975.2023.2229544](https://doi.org/10.1080/23311975.2023.2229544).
- [3] "Digital Wallet Research 2026: User Behavior & Competitive Landscape," Jakarta, 2026. [Online]. Available: <https://www.ipsos.com/en-id/mapping-digital-wallet-landscape-2026-which-platform-leads-users-preferred-choice-according-ipsos>
- [4] J. Yi, J. Kim, and Y. K. Oh, "Uncovering the quality factors driving the success of mobile payment apps," *J. Retail. Consum. Serv.*, vol. 77, pp. 1–9, 2024, doi: [10.1016/j.jretconser.2023.103641](https://doi.org/10.1016/j.jretconser.2023.103641).
- [5] P. A. Aritonang, M. E. Johan, and I. Prasetyawan, "Aspect-Based Sentiment Analysis on Application Review using CNN," *Ultim. InfoSys J. Ilmu Sist. Inf.*, vol. 13, no. 1, pp. 54–61, 2022, doi: [10.31937/si.v13i1.2684](https://doi.org/10.31937/si.v13i1.2684).
- [6] E. Yulianti and N. K. Nissa, "ABSA of Indonesian customer reviews using IndoBERT: single-sentence and sentence-pair classification approaches," *Bull. Electr. Eng. Informatics*, vol. 13, no. 5, pp. 3579–3589, 2024, doi: [10.11591/eei.v13i5.8032](https://doi.org/10.11591/eei.v13i5.8032).
- [7] A. R. Aisy and G. Karyono, "Analisis Sentimen Terhadap Ulasan Google Play Store Aplikasi Lazada, Shopee, dan Tokopedia Menggunakan Algoritma IndoBERT," *Build. Informatics, Technol. Sci.*, vol. 7, no. 3, pp. 2127–2135, 2025, doi: [10.47065/bits.v7i3.8745](https://doi.org/10.47065/bits.v7i3.8745).
- [8] T. N. Prakash and A. Aloysius, "Textual Sentiment Analysis using Lexicon Based Approaches," *Ann. Rom. Soc. Cell Biol.*, vol. 25, no. 4, pp. 9878–9885, 2021.
- [9] P. Shukla, R. Kumar, V. K. Dwivedi, and A. K. Singh, "Aspect based sentiment analysis: A systematic review, taxonomy, applications, and future research directions," *Comput. Sci. Rev.*, vol. 61, p. 100924, 2026, doi: [10.1016/j.cosrev.2026.100924](https://doi.org/10.1016/j.cosrev.2026.100924).
- [10] Y. Li *et al.*, "Better Queries for Aspect-Category Sentiment Classification," in *Proceedings of the 19th China National Conference on Computational Linguistics (CCL 2020)*, Hainan, China: Technical Committee on Computational Linguistics, Chinese Information Processing Society of China, 2020, pp. 1079–1088.
- [11] U. Fayyad and R. Uthurusamy, "Data Mining and Knowledge Discovery in Databases," *Commun. ACM*, vol. 39, no. 11, pp. 24–26, 1996, doi: [10.1145/240455.240463](https://doi.org/10.1145/240455.240463).
- [12] A. Fatihaturrahmah and K. D. Tania, "Review: A Hybrid Approach of Aspect-Based Sentiment Analysis and Knowledge Extraction for Evaluating Security Perceptions in Digital Payment Applications," *Sci. J. Informatics*, vol. 12, no. 4, pp. 635–650, 2025, doi: [10.15294/sji.v12i4.31557](https://doi.org/10.15294/sji.v12i4.31557).
- [13] S. A. Putri, K. D. Tania, and Naretha Kawadha Pasemah Gumay, "Knowledge Discovery Through Sentiment Analysis and Topic Modeling of BCA Mobile and MyBCA," *J. Math. Comput. Stat.*, vol. 8, no. 2, pp. 669–682, 2025, doi: [10.35580/jmathcos.v8i2.9782](https://doi.org/10.35580/jmathcos.v8i2.9782).
- [14] M. Nashiroh Ramadhani, K. Ditha Tania, and M. Afrina, "Knowledge Discovery in Sharia Mobile Banking Reviews Using Aspect-Based Sentiment Analysis and Machine Learning," *J. Appl. Informatics Comput.*, vol. 10, no. 1, pp. 640–650, 2026, doi: [10.35580/jmathcos.v8i2.9782](https://doi.org/10.35580/jmathcos.v8i2.9782).

- 10.30871/jaic.v10i1.11753.
- [15] H. Jayadianti, W. Kaswidjanti, A. T. Utomo, S. Saifullah, F. A. Dwiyanto, and R. Drezewski, "Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN," *Ilk. J. Ilm.*, vol. 14, no. 3, pp. 348–354, 2022, doi: 10.33096/ilkom.v14i3.1505.348-354.
 - [16] H. Imaduddin, F. Y. A'la, and Y. S. Nugroho, "Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 8, pp. 113–117, 2023, doi: 10.14569/IJACSA.2023.0140813.
 - [17] Sandy Sanjaya, Rangga Gelar Guntara, and Syti Sarah Maesaroh, "Sentiment Analysis of LinkAja Digital Wallet Application Reviews on Google Play Store using Transfer Learning IndoBERT," *INOVTEK Polbeng - Seri Inform.*, vol. 10, no. 3, pp. 1730–1740, 2025, doi: 10.35314/afjx7b16.
 - [18] I. Daqilq, H. Saputra, Syamsudhuha, R. Kurniawan, and Y. Andriyani, "Sentiment analysis of student evaluation feedback using transformer-based language models," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 36, no. 2, pp. 1127–1139, 2024, doi: 10.11591/ijeecs.v36.i2.pp1127-1139.
 - [19] Fransiscus and A. S. Girsang, "Sentiment Analysis of COVID-19 Public Activity Restriction (PPKM) Impact using BERT Method," *Int. J. Eng. Trends Technol.*, vol. 70, no. 12, pp. 281–288, 2022, doi: 10.14445/22315381/IJETT-V70I12P226.
 - [20] N. S. Desvira and M. F. Aransyah, "Analisis Faktor-Faktor yang Memengaruhi Minat dan Perilaku Penggunaan Fitur ShopeePay Menggunakan Model Unified Theory of Acceptance and Use of Technology (UTAUT2)," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 12, no. 2, pp. 178–191, 2023, doi: <https://doi.org/10.32736/sisfokom.v12i2.1594>.
 - [21] H. Nur Alfiana, A. Doewes, and B. Widoyono, "Aspect-Based Sentiment Analysis of Access by KAI Application Reviews Using IndoBERT for Multi-Label Classification Tasks," *J. Tek. Inform.*, vol. 7, no. 1, pp. 286–306, 2026, doi: 10.52436/1.jutif.2026.7.1.5402.
 - [22] F. A. D. Aryanti, A. Luthfiarta, and D. A. I. Soeroso, "Aspect-Based Sentiment Analysis with LDA and IndoBERT Algorithm on Mental Health App: Riliv," *J. Appl. Informatics Comput.*, vol. 9, no. 2, pp. 361–375, 2025, doi: 10.30871/jaic.v9i2.8958.
 - [23] A. P. Maretta and A. Meiriza, "Aspect-Based Sentiment Analysis of Hospital Service Reviews Using Fine-Tuned IndoBERT," *J. Appl. Informatics Comput.*, vol. 9, no. 5, pp. 2541–2551, 2025, doi: 10.30871/jaic.v9i5.10765.
 - [24] K. S. Nugroho, A. Y. Sukmadewa, H. Wuswilahaken Dw, F. A. Bachtiar, and N. Yudistira, "BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews," *ACM Int. Conf. Proceeding Ser.*, pp. 258–264, 2021, doi: 10.1145/3479645.3479679.
 - [25] U. Rahardja, C. T. Sigalingging, P. O. H. Putra, A. N. Hidayanto, and K. Phusavat, "The Impact of Mobile Payment Application Design and Performance Attributes on Consumer Emotions and Continuance Intention," *SAGE Open*, vol. 13, no. 1, pp. 1–18, 2023, doi: 10.1177/21582440231151919.
 - [26] A. Valencia-Arias *et al.*, "Research trends in mobile payment adoption: Research trends and agenda," *F1000Research*, vol. 14, pp. 1–35, 2025, doi: 10.12688/f1000research.159551.2.
 - [27] M. A. Sait, M. Anshari, and M. N. Almunawar, "Understanding factors to digital wallet discontinuance intention among past users: an exploratory study," *J. Sci. Technol. Policy Manag.*, vol. 17, no. 2, pp. 212–231, 2024, doi: 10.1108/JSTPM-01-2024-0005.
 - [28] T. Zhou, "An empirical examination of continuance intention of mobile payment services," *Decis. Support Syst.*, vol. 54, no. 2, pp. 1085–1091, 2013, doi: 10.1016/j.dss.2012.10.034.
 - [29] A. Kumar, P. Haldar, and S. Chaturvedi, "Factors influencing intention to continue use of e-wallet: mediating role of perceived usefulness," *Vilakshan - XIMB J. Manag.*, vol. 22, no. 1, pp. 45–61, 2025, doi: 10.1108/xjm-12-2023-0243.
 - [30] B. Liu, "Sentiment Analysis and Opinion Mining," *Synth. Lect. Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, May 2012, doi: 10.2200/S00416ED1V01Y201204HLT016.
 - [31] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960, doi: <https://doi.org/10.1177/00131644600200010>.
 - [32] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977, doi: <https://doi.org/10.2307/2529310>.
 - [33] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2, pp. 1137–1143, 1995.
 - [34] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, 2nd ed., vol. 102. in Springer Texts in Statistics, vol. 102. New York, NY: Springer US, 2021. doi: 10.1007/978-1-0716-1418-1.
 - [35] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," *ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 2623–2631, 2019, doi: 10.1145/3292500.3330701.
 - [36] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 4171–4186, 2019.
 - [37] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
 - [38] P. Ayuningtiyas, K. D. Tania, and W. K. Sari, "Sentiment-Based Knowledge Discovery pada Aplikasi iPusnas Menggunakan Metode Machine Learning dan Deep Learning," *J. Appl. Informatics Comput.*, vol. 9, no. 5, pp. 2486–2497, Oct. 2025, doi: 10.30871/jaic.v9i5.10258.
 - [39] M. R. Maulana, K. D. Tania, and P. Sari, "Comparative Analysis of Machine Learning Models for Daily Urban Waste Volume's Knowledge Discovery," *Int. Conf. Artif. Intell. Technol. Solut.*, pp. 131–136, 2025, doi: 10.1109/ICAITEch66481.2025.11387653.
 - [40] Z. Tri Zafira, K. Ditha Tania, and W. Kurnia Sari, "Sentiment-Based Knowledge Discovery of Wondr by BNI App Reviews Using SVM, KNN, and Naive Bayes for CRM Enhancement," *J. Appl. Informatics Comput.*, vol. 9, no. 5, pp. 2498–2508, Oct. 2025, doi: 10.30871/jaic.v9i5.10323.