Analisis Sentimen Multi-Aspek terhadap Ulasan Pengguna Aplikasi M-Pajak Menggunakan Model IndoBERT

Multi-Aspect Sentiment Analysis of M-Pajak Application User Reviews Using the IndoBERT Model

Hansen Utomo Gunawan[1], **Fergie Joanda Kaunang***[2]
Program Studi Informatika, Fakultas Teknologi dan Desain, Universitas Bunda Mulia
E-mail : hansenutomoo@gmail.com[1], fkaunang@bundamulia.ac.id[2]*
**Corresponding author*
Received 30 January 2026; Revised 9 February 2026; Accepted 22 February 2026

Abstract - Aplikasi M-Pajak menghadapi tantangan dalam evaluasi layanan akibat banyaknya ulasan tidak terstruktur. Penelitian ini bertujuan untuk menganalisis sentimen multi-aspek menggunakan model *Deep Learning* berbasis arsitektur IndoBERT yang dioptimalkan dengan teknik *Random Oversampling* untuk menangani ketidakseimbangan data (*class imbalance*). Lima aspek dominan diekstraksi menggunakan metode *Latent Dirichlet Allocation* (LDA) yang divalidasi dengan nilai *Topic Coherence*. Hasil pengujian menunjukkan bahwa model mampu mencapai rata-rata *Macro-F1-Score* sebesar 93.86% pada kelima aspek yang diuji. Validitas kinerja model diperkuat oleh analisis kurva *Receiver Operating Characteristic* (ROC) yang mencatatkan nilai *Area Under Curve* (AUC) rata-rata di atas 0.99, serta pengujian 10-*Fold Cross-Validation* yang menghasilkan standar deviasi rendah (±0,0032), mengindikasikan bahwa model memiliki daya diskriminasi yang sangat tinggi dan tegas dalam membedakan sentimen negatif, netral, dan positif. Penelitian ini menyimpulkan bahwa integrasi IndoBERT dengan penanganan data tidak seimbang sangat efektif untuk klasifikasi teks ulasan layanan publik yang kompleks.

Kata kunci - Analisis sentimen, IndoBERT, M-Pajak, multi-aspek, *Latent Dirichlet Allocation*.

Abstract - *The M-Pajak application faces challenges in service evaluation due to the large volume of unstructured reviews. This study aims to perform multi-aspect sentiment analysis using the IndoBERT Deep Learning model architecture, optimized with the Random Oversampling technique to address class imbalance issues. Five dominant aspects were extracted using Latent Dirichlet Allocation (LDA) and validated through topic coherence analysis. The results show that the model achieved an average Macro-F1-Score of 93.86% across the five tested aspects. The model's validity is further confirmed by Receiver Operating Characteristic (ROC) curve analysis, recording an average Area Under Curve (AUC) above 0.99, and 10-Fold Cross-Validation, which yielded a low standard deviation (±0.0032), indicating exceptional discriminatory power in distinguishing between negative, neutral, and positive sentiments. This study concludes that integrating IndoBERT with imbalanced data handling is highly effective for classifying complex public service review texts.*

Keywords - *Sentiment analysis, IndoBERT, M-Pajak, multi-aspect, Latent Dirichlet Allocation.*

1. PENDAHULUAN

Transformasi digital dalam sektor perpajakan di Indonesia telah melahirkan inovasi berupa aplikasi M-Pajak yang dirancang untuk memudahkan wajib pajak dalam mengakses layanan seperti pembuatan kode *billing* dan konfirmasi status wajib pajak [1]. Meskipun bertujuan meningkatkan efisiensi, aplikasi ini menghadapi tantangan signifikan terkait kepuasan pengguna. Berdasarkan data ulasan di Google Play Store per Oktober 2025, M-Pajak menerima mayoritas rating rendah dengan rata-rata 2.8 bintang, di mana ribuan ulasan tersebut bersifat tidak terstruktur. Evaluasi terhadap ulasan ini menjadi krusial karena opini pengguna mengandung

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15761

wawasan mendalam mengenai aspek spesifik layanan yang perlu diperbaiki, mulai dari sisi teknis hingga kualitas pelayanan [2], [18].

Tantangan utama dalam menganalisis ulasan pengguna secara manual adalah volume data yang besar dan format teks yang tidak baku [3]. Oleh karena itu, pendekatan analisis sentimen berbasis aspek (*aspect-based sentiment analysis*) menjadi solusi yang diperlukan untuk memetakan permasalahan secara granular [4]. Beberapa penelitian terdahulu telah mencoba menganalisis sentimen pada layanan publik serupa menggunakan algoritma machine learning tradisional . Penelitian sebelumnya oleh Wijaya menerapkan Naive Bayes pada aplikasi SIGNAL dengan akurasi 63,61% [5]. Sementara penelitian oleh Faisol membandingkan *Support Vector Machine* (SVM), Naive Bayes, dan *K-Nearest Neighbors* (KNN) pada aplikasi M-Pajak, di mana SVM menghasilkan kinerja terbaik [6]. Namun, metode tradisional memiliki keterbatasan dalam memahami konteks linguistik bahasa Indonesia yang kompleks dan sering kali gagal menangkap nuansa kalimat yang memiliki dependensi jangka panjang .

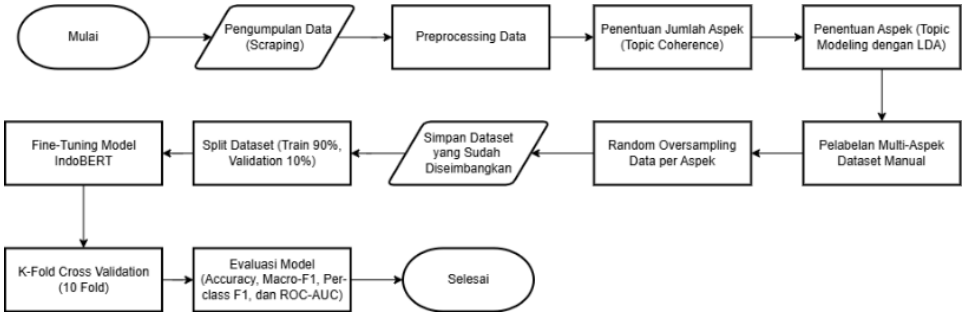
Untuk mengatasi keterbatasan tersebut, pendekatan *deep learning* berbasis arsitektur *transformer*, khususnya IndoBERT, menawarkan solusi yang lebih mutakhir. Studi internasional oleh Jahan et al. ditahun 2024 terhadap deteksi ujaran kebencian juga menunjukkan bahwa model berbasis *transformer*, termasuk BERT dan RoBERTa, mampu meningkatkan *accuracy* hingga 5–12% dibandingkan pendekatan tradisional seperti *Long Short-Term Memory* (LSTM) atau SVM [7]. IndoBERT merupakan model *pre-trained* yang dilatih pada korpus bahasa Indonesia berskala besar, sehingga memiliki representasi kontekstual yang lebih kaya dibandingkan model statistik biasa [8]. Meskipun IndoBERT telah banyak diterapkan, tantangan spesifik dalam data ulasan aplikasi adalah ketidakseimbangan kelas (*class imbalance*), di mana jumlah ulasan netral sering kali mendominasi dibandingkan ulasan negatif atau positif [9]. Ketimpangan ini dapat menyebabkan model menjadi bias dan memiliki performa rendah pada kelas minoritas [10]. Menilai sentimen secara umum dari setiap ulasan saja tidaklah cukup, karena pengguna sering membahas berbagai aspek dari aplikasi atau topik, seperti tampilan, kecepatan, stabilitas, kemudahan, hingga kualitas layanan [11].

Penelitian ini bertujuan untuk mengevaluasi kinerja model IndoBERT dalam melakukan analisis sentimen multi-aspek terhadap ulasan M-Pajak. Berbeda dengan penelitian sebelumnya yang hanya berfokus pada klasifikasi sentimen umum, penelitian ini menerapkan strategi *fine-tuning* pada lima aspek spesifik (autentikasi data, verifikasi data, pengalaman pengguna, *usability system*, dan layanan pengguna). Fokus utama penelitian ini terletak pada optimasi arsitektur model melalui teknik *Random Oversampling* untuk menangani *imbalanced data* serta validasi kinerja model menggunakan metriks evaluasi. Meskipun penelitian ini pada akhirnya bertujuan untuk menyediakan landasan analitik bagi pengembangan purwarupa aplikasi *dashboard* pemantauan, penelitian ini akan memusatkan pembahasan pada eksperimen komputasi, konfigurasi *hyperparameter*, dan analisis metrik evaluasi model (*accuracy, precision, recall, F1-score*) serta validasi daya diskriminasi model menggunakan kurva *Receiver Operating Characteristic-Area Under the Curve* (ROC-AUC).

2. METODE PENELITIAN

Tahapan penelitian dimulai dari pengumpulan data melalui *web scraping*, ekstraksi lima aspek menggunakan metode LDA, dan pelabelan manual. Setelah tahap pra-pemrosesan, dilakukan teknik *Random Oversampling* untuk menyeimbangkan data pada tiap aspek. Selanjutnya, dilakukan *fine-tuning* model IndoBERT secara iteratif dengan pembagian data 90% *train* dan 10% *validation* hingga diperoleh model terbaik. Seluruh rangkaian proses ditutup dengan evaluasi performa menggunakan metrik *F1-score* dan AUC untuk menjamin akurasi klasifikasi sentimen. Alur lengkap penelitian ini dapat dilihat pada Gambar 1.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15761



Gambar 1. *Flowchart* Penelitian

2.1. Pengumpulan Data

Data penelitian berupa ulasan pengguna aplikasi M-Pajak yang diperoleh dari *platform* Google Play Store. Proses pengumpulan data dilakukan menggunakan teknik *web scraping*. Filter diterapkan untuk mengambil ulasan berbahasa Indonesia dengan rentang waktu dari Juni 2021 hingga Oktober 2025. Total data mentah yang berhasil dikumpulkan sebanyak 5.709 ulasan, yang mencakup atribut teks ulasan, rating bintang, dan tanggal ulasan.

2.2. Pra-pemrosesan Data (*Preprocessing*)

Untuk memastikan kualitas data masukan model, dilakukan serangkaian tahapan pra-pemrosesan teks. Tahapan pra-pemrosesan data dilakukan secara komprehensif untuk menjamin kualitas teks sebelum memasuki fase pemodelan. Proses ini dimulai dengan pembersihan data (*cleaning*) menggunakan *Regular Expression* (Regex) untuk menghapus karakter non-alfanumerik, URL, *mentions*, serta tanda baca yang dianggap tidak relevan bagi analisis [12], [19]. Selanjutnya, dilakukan normalisasi kata guna mengubah kosakata tidak baku atau bahasa gaul menjadi bentuk baku berdasarkan kamus normalisasi yang telah ditentukan, namun tetap memberikan pengecualian pada kata kunci sentimen krusial seperti kata "ok" atau "oke" agar bobot maknanya tetap terjaga. Sebagai tahap akhir, diterapkan penghapusan *stopwords* untuk menghilangkan kata-kata hubung umum yang tidak memiliki signifikansi terhadap sentimen guna mengurangi tingkat gangguan (*noise*), dengan tetap mempertahankan kata-kata negasi yang dianggap penting untuk menjaga konteks sentimen dalam ulasan [13].

Tabel 1. Hasil *Pre-processing*

Ulasan Asli	Hasil *Pre-processing*
Sy sgt kecewa karena pelayanan di kantor DJP Serang sangat buruk.	saya sangat kecewa pelayanan kantor djp serang sangat buruk
Apl berjalan dgn baik, sukses apl m-pajak.	aplikasi berjalan baik sukses aplikasi mpajak

2.3. Penentuan Aspek dan Pelabelan

Penelitian ini tidak menggunakan aspek yang ditentukan secara apriori, melainkan menemukannya secara induktif dari data menggunakan *Latent Dirichlet Allocation* (LDA). LDA digunakan untuk memodelkan topik dominan dari korpus ulasan [14]. Dilakukan *Topic Coherence* yang merupakan metrik evaluasi untuk mengukur keterkaitan semantik dan konsistensi kata-kata dalam sebuah topik yang dihasilkan oleh model LDA [17]. Skor koherensi digunakan untuk memastikan bahwa topik-topik tersebut bermakna dan mencerminkan struktur informasi asli dari dataset. Semakin tinggi nilai koherensi, semakin kuat hubungan semantik antar kata kunci, yang mengindikasikan bahwa model tersebut lebih andal dan dapat dipercaya dalam menjelaskan struktur topikal data. Hasil skor koherensi untuk penelitian ini dapat dilihat pada Tabel 2.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15761

Tabel 2. Hasil Pengujian Skor *Topic Coherence* (C*v*)

Jumlah Topik (K)	***Coherence Score* (C*v*)**
2	0.6153
3	0.5936
4	0.5788
5	0.5971
6	0.5923
7	0.5857
8	0.5950
9	0.5758
10	0.5771

Meskipun nilai K=2 mencatatkan skor tertinggi (0.6153), jumlah tersebut dinilai terlalu luas (*broad*) untuk kebutuhan analisis multi-aspek yang mendalam pada ulasan aplikasi. Pemilihan K=5 didasarkan pada adanya puncak lokal (*local peak*) sebesar 0,5971 yang lebih tinggi dibandingkan jumlah topik lainnya. Berdasarkan nilai probabilitas kata kunci tertinggi, diidentifikasi lima aspek utama layanan M-Pajak: (1) Autentikasi Data, (2) Verifikasi Data, (3) Pengalaman Pengguna, (4) *Usability System*, dan (5) Layanan Pengguna. Setelah aspek ditentukan, dilakukan pelabelan data secara manual oleh peneliti untuk menentukan sentimen (positif, negatif, netral) pada setiap aspek di masing-masing ulasan. Dataset berlabel ini kemudian dibagi menjadi data latih (*training*) dan data validasi.

Tabel 3. Hasil *Topic Modeling* dengan LDA

Hasil *Topic Modeling*
0.082*"login" + 0.072*"npwp" + 0.057*"daftar" + 0.040*"benar" + 0.036*"padahal" + 0.036*"salah" + 0.032*"malah" + 0.026*"mantap" + 0.026*"web" + 0.020*"password"
0.057*"baik" + 0.050*"terus" + 0.040*"gagal" + 0.037*"data" + 0.032*"guna" + 0.026*"lupa" + 0.025*"selalu" + 0.021*"semua" + 0.021*"isi" + 0.021*"error"
0.063*"masuk" + 0.060*"efin" + 0.041*"verifikasi" + 0.040*"kode" + 0.035*"email" + 0.029*"nomor" + 0.026*"minta" + 0.026*"coba" + 0.024*"kali" + 0.021*"kirim"
0.078*"buat" + 0.070*"mau" + 0.044*"susah" + 0.041*"mudah" + 0.040*"bayar" + 0.039*"lapor" + 0.036*"bagus" + 0.034*"kalau" + 0.030*"apa" + 0.030*"sulit"
0.097*"sangat" + 0.048*"rumit" + 0.046*"bantu" + 0.040*"sama" + 0.030*"kantor" + 0.028*"spt" + 0.024*"perintah" + 0.022*"sekali" + 0.021*"jelas" + 0.020*"layan"

Proses interpretasi kata kunci menjadi aspek dilakukan dengan mengelompokkan kata-kata dengan probabilitas tertinggi yang memiliki kesamaan domain layanan, sebagaimana dirinci pada Tabel 4.

Tabel 4. Pemetaan Topik LDA ke Aspek Layanan

Label Aspek	**Kata Kunci Dominan**	**Justifikasi Pemetaan**
Autentikasi Data	login, npwp, daftar, benar	Berkaitan dengan akses masuk dan pendaftaran akun.
Verifikasi Data	efin, verifikasi, kode, otp, email	Berkaitan dengan validasi/verifikasi keamanan dan pengiriman kode.
Pengalaman Pengguna	buat, mau, susah, mudah, bagus	Merujuk pada impresi subjektif alur penggunaan aplikasi.
Usability System	gagal, *error*, baik, buka, isi	Berkaitan dengan performa teknis dan stabilitas sistem.
Layanan Pengguna	bantu, layan, kantor, jelas	Berkaitan dengan fungsi administratif dan operasional.

2.4. Penanganan Ketidakseimbangan Data (*Data Balancing*)

Analisis awal menunjukkan adanya ketidakseimbangan distribusi kelas yang signifikan (*imbalanced data*), di mana sentimen netral mendominasi mayoritas ulasan. Untuk mencegah bias

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15761

model pada kelas mayoritas, diterapkan teknik *Random Oversampling* [15]. Teknik ini menduplikasi sampel dari kelas minoritas (positif dan negatif) secara acak hingga proporsinya setara dengan kelas mayoritas pada setiap aspek. Hal ini menghasilkan dataset pelatihan yang seimbang untuk mengoptimalkan proses pembelajaran model.

2.5. Arsitektur Model dan Pelatihan

Model klasifikasi dibangun menggunakan arsitektur IndoBERT (IndoBERT-base-p2), sebuah model *pre-trained* berbasis *transformer* yang telah dilatih pada korpus Bahasa Indonesia berskala besar [8]. Penelitian ini mengusulkan strategi *One-Model-Per-Aspect*, di mana lima model IndoBERT independen dilatih (*fine-tuned*) secara terpisah untuk setiap aspek. Pendekatan ini dipilih agar setiap model dapat mempelajari pola linguistik spesifik yang relevan dengan aspek masing-masing (misalnya, istilah pada aspek 'Autentikasi' berbeda konteks dengan 'Layanan'). Konfigurasi *hyperparameter tuning* dilakukan menggunakan metode *grid search* untuk mendapatkan performa optimal.

2.6. Evaluasi Model

Kinerja model dievaluasi menggunakan metrik evaluasi yang meliputi *accuracy*, *precision*, *recall*, dan *F1-score*, yang dihitung berdasarkan *confusion matrix*. Fokus utama evaluasi adalah nilai *F1-score* karena kemampuannya memberikan gambaran performa yang berimbang antara *precision* dan *recall* pada kasus klasifikasi multi-kelas [11]. *Accuracy* merupakan perbandingan antara jumlah prediksi yang tepat, baik positif maupun negatif, dengan keseluruhan jumlah data [16], [20].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

Keterangan:
TP: *True Positive*, jumlah data positif yang diprediksi benar.
TN: *True Negative*, jumlah data negatif yang diprediksi benar.
FP: *False Positive*, jumlah data negatif yang diprediksi salah sebagai positif.
FN: *False Negative*, jumlah data positif yang diprediksi salah sebagai negatif.

$$Precision = \frac{TP}{TP+FP} \qquad (2)$$

Precision mengukur proporsi prediksi positif yang benar (TP) dibandingkan dengan seluruh prediksi positif yang dihasilkan oleh model [16].

$$Recall = \frac{TP}{TP+FN} \qquad (3)$$

Recall menunjukkan kemampuan model dalam menemukan seluruh data positif yang benar, dihitung sebagai rasio antara TP dengan jumlah data positif sebenarnya [16].

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (4)$$

F1-score juga dikenal sebagai *F-measure* adalah rata-rata harmonik antara *precision* dan *recall*, yang digunakan untuk menyeimbangkan keduanya, terutama ketika terdapat ketidakseimbangan kelas [16].

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15761

3. HASIL DAN PEMBAHASAN

Hasil eksperimen komputasi yang dilakukan untuk mengevaluasi kinerja model IndoBERT dalam mengklasifikasikan sentimen multi-aspek mencakup pencarian parameter optimal (*hyperparameter tuning*), serta pengujian akhir pada dataset yang telah diseimbangkan. Penelitian ini menerapkan metode *grid search* dengan menguji kombinasi *Epoch* (2, 3, 4), *Batch Size* (16, 32), dan *Learning Rate* (1e-5 hingga 5e-5). Hasil pengujian menunjukkan bahwa peningkatan *Learning Rate* ke angka 5e-5 cenderung memberikan stabilitas yang lebih baik. Kombinasi parameter terbaik tercapai pada *Epoch* 2, *Batch Size* 32, dan *Learning Rate* 5e-5, yang menghasilkan rata-rata *F1-score* tertinggi. Rangkuman hasil *tuning* disajikan pada Tabel 3.

Tabel 5. Hasil Eksperimen *Hyperparameter Tuning*

Epoch	*Batch*	LR	Autentikasi		Verifikasi		Pengalaman		*Usability*		Layanan	
			F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
2	16	1e-5	0.81	0.86	0.81	0.87	0.84	0.84	0.79	0.82	0.80	0.90
		2e-5	0.81	0.86	0.82	0.87	0.84	0.84	0.80	0.82	0.80	0.90
		3e-5	0.81	0.86	0.82	0.87	0.83	0.84	0.80	0.82	0.81	0.90
		4e-5	0.81	0.86	0.82	0.87	0.83	0.84	0.80	0.82	0.81	0.91
		5e-5	0.81	0.86	0.82	0.87	0.83	0.84	0.80	0.82	0.81	0.91
	32	1e-5	0.80	0.86	0.81	0.87	0.82	0.84	0.79	0.81	0.79	0.90
		2e-5	0.80	0.86	0.83	0.88	0.82	0.84	0.81	0.83	0.80	0.90
		3e-5	0.80	0.86	0.83	0.88	0.82	0.84	0.81	0.82	0.81	0.90
		4e-5	0.80	0.85	0.84	0.89	0.83	0.83	0.82	0.82	0.81	0.90
		5e-5	**0.80**	**0.85**	**0.84**	**0.89**	**0.83**	**0.83**	**0.82**	**0.82**	**0.81**	**0.90**
3	16	1e-5	0.81	0.86	0.82	0.87	0.83	0.84	0.82	0.84	0.80	0.89
		2e-5	0.81	0.86	0.82	0.87	0.83	0.84	0.80	0.82	0.80	0.89
		3e-5	0.81	0.86	0.83	0.87	0.82	0.84	0.80	0.82	0.80	0.89
		4e-5	0.81	0.86	0.83	0.87	0.82	0.83	0.80	0.82	0.80	0.89
		5e-5	0.81	0.86	0.83	0.87	0.82	0.83	0.80	0.82	0.80	0.89
	32	1e-5	0.81	0.86	0.82	0.87	0.82	0.83	0.79	0.82	0.79	0.90
		2e-5	0.81	0.86	0.83	0.87	0.83	0.84	0.80	0.82	0.80	0.90
		3e-5	0.81	0.86	0.83	0.88	0.83	0.84	0.80	0.82	0.80	0.90
		4e-5	0.81	0.85	0.83	0.88	0.83	0.84	0.80	0.83	0.80	0.90
		5e-5	0.81	0.85	0.83	0.88	0.83	0.84	0.80	0.83	0.80	0.90
4	16	1e-5	0.81	0.86	0.82	0.87	0.82	0.83	0.80	0.82	0.80	0.89
		2e-5	0.81	0.86	0.82	0.87	0.83	0.84	0.80	0.82	0.80	0.90
		3e-5	0.81	0.86	0.82	0.87	0.83	0.84	0.80	0.82	0.80	0.90
		4e-5	0.81	0.86	0.82	0.87	0.83	0.84	0.80	0.82	0.80	0.90
		5e-5	0.81	0.86	0.83	0.87	0.83	0.84	0.80	0.82	0.80	0.90
	32	1e-5	0.81	0.86	0.81	0.87	0.82	0.83	0.79	0.81	0.79	0.89
		2e-5	0.81	0.86	0.82	0.87	0.82	0.83	0.80	0.82	0.80	0.90
		3e-5	0.81	0.86	0.83	0.87	0.82	0.84	0.80	0.82	0.80	0.90
		4e-5	0.81	0.86	0.83	0.88	0.83	0.84	0.80	0.83	0.80	0.90
		5e-5	0.81	0.86	0.83	0.88	0.83	0.84	0.80	0.83	0.80	0.90

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15761

Evaluasi model dilakukan melalui dua skenario utama untuk memastikan objektivitas hasil: validasi silang (*cross-validation*) pada data latih untuk mengukur stabilitas arsitektur, dan pengujian akhir pada data yang telah melalui proses penyeimbangan (*balancing*). Penerapan strategi 10-*Fold Cross-Validation* pada dataset asli (*imbalanced*) menghasilkan rata-rata akurasi sebesar 83,97% dan *F1-score* sebesar 79,31% dengan standar deviasi ±0,0177. Angka variansi yang cukup tinggi ini mengonfirmasi adanya fluktuasi performa akibat dominasi kelas 'netral' yang menghambat model dalam mengenali fitur pada kelas minoritas.

Namun, setelah penerapan teknik *Random Oversampling* pada skenario kedua, performa model meningkat signifikan dengan rata-rata akurasi 89,86% dan *F1-score* 89,80%. Selain peningkatan akurasi, tingkat stabilitas model juga meningkat drastis yang dibuktikan dengan penurunan standar deviasi menjadi ±0,0036. Hal ini menunjukkan bahwa penyeimbangan data tidak hanya meningkatkan ketepatan prediksi, tetapi juga membuat model IndoBERT menjadi lebih konsisten (*robust*) dalam melakukan klasifikasi sentimen di seluruh bagian dataset. Hasil pengujian kinerja model dengan penerapan 10-*Fold Cross-Validation* dapat dilihat pada Tabel 6 dan Tabel 7.

Tabel 6. Evaluasi Kinerja Model Per *Fold* Menggunakan Dataset *Imbalance* pada *Single-Model*

Fold	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Fold 1	0.8395	0.8375	0.7754	0.8010
Fold 2	0.8354	0.7896	0.7859	0.7873
Fold 3	0.8287	0.7985	0.7814	0.7886
Fold 4	0.8563	0.8257	0.7967	0.8101
***Fold* 5**	**0.8692**	**0.8350**	**0.8241**	**0.8277**
Fold 6	0.8288	0.7912	0.7609	0.7716
Fold 7	0.8421	0.8106	0.7962	0.8022
Fold 8	0.8246	0.7935	0.7619	0.7746
Fold 9	0.8397	0.8058	0.7830	0.7935
Fold 10	0.8329	0.7914	0.7652	0.7746
Rata-Rata	**0.8397**	**0.8078**	**0.7830**	**0.7931**
Standar Deviasi (±)	**0.0136**	**0.0186**	**0.0193**	**0.0177**

Tabel 7. Evaluasi Kinerja Model Per *Fold* Menggunakan Dataset *Balance* pada *Single-Model*

Fold	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Fold 1	0.8978	0.8973	0.8972	0.8966
Fold 2	0.8997	0.9	0.8998	0.8992
Fold 3	0.8967	0.8966	0.8973	0.8962
Fold 4	0.8975	0.8998	0.8976	0.8975
***Fold* 5**	**0.9063**	**0.9074**	**0.9057**	**0.9054**
Fold 6	0.8967	0.8985	0.898	0.897
Fold 7	0.8956	0.8958	0.8954	0.8948
Fold 8	0.9037	0.905	0.9041	0.9033
Fold 9	0.8961	0.8968	0.896	0.8957
Fold 10	0.8959	0.8976	0.8948	0.8943
Rata-rata	**0.8986**	**0.8994**	**0.8985**	**0.8980**
Standar Deviasi (±)	**0.0036**	**0.0038**	**0.0036**	**0.0036**

Meskipun pendekatan *Single-Model-Multi-Aspect* mampu memberikan hasil performa yang cukup baik secara global, penelitian ini memilih untuk menerapkan pendekatan *One-Model-*

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15761

Per-Aspect dengan beberapa alasan teknis yang mendasar. Setiap aspek dalam ulasan M-Pajak (seperti 'Autentikasi Data' dan *'Usability System*') memiliki karakteristik linguistik dan terminologi yang sangat spesifik. Dengan melatih satu model khusus untuk setiap aspek, arsitektur IndoBERT dapat melakukan *fine-tuning* yang lebih mendalam terhadap kosakata unik pada domain aspek tersebut tanpa terdistraksi oleh pola bahasa dari aspek lain. Hal ini terbukti mampu meminimalkan tingkat ambiguitas klasifikasi pada kalimat yang mengandung keluhan teknis yang serupa namun merujuk pada aspek yang berbeda.

Pendekatan *One-Model-Per-Aspect* juga memungkinkan optimasi dataset yang lebih presisi melalui teknik *Random Oversampling* pada setiap kategori secara independen. Pada pendekatan *Single-Model-Multi-Aspect*, penyeimbangan data seringkali sulit dilakukan secara merata karena satu ulasan bisa mengandung label sentimen yang berbeda untuk aspek yang berbeda (*multi-label*). Dengan memisahkan model, setiap aspek mendapatkan proporsi distribusi kelas yang benar-benar seimbang, sebagaimana ditunjukkan pada Tabel 9. Hal ini divalidasi oleh hasil eksperimen yang menunjukkan peningkatan rata-rata *F1-Score* sebesar 14,56% jika membandingkan hasil pada Tabel 8 (*Dataset Imbalance*) dengan Tabel 9 (*Dataset Balance*). Penggunaan strategi ini menjamin bahwa sistem analisis sentimen yang dibangun memiliki daya diskriminasi yang lebih *robust* dan tajam dalam menangani karakteristik ulasan aplikasi layanan publik yang kompleks.

Tabel 8. Evaluasi Kinerja Model Menggunakan Pendekatan *One-Model-Per-Aspect* dengan Dataset *Imbalance*

Aspek	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Autentikasi Data	0.8607	0.8238	0.8194	0.8214
Verifikasi Data	0.8794	0.8616	0.8335	0.8462
Pengalaman Pengguna	0.8295	0.8238	0.8202	0.8219
Usability System	0.8274	0.8182	0.8118	0.8147
Layanan Pengguna	0.8877	0.8287	0.7723	0.7927
Rata-Rata	**0.8569**	**0.8312**	**0.8114**	**0.8193**

Tabel 9. Evaluasi Kinerja Model Menggunakan Pendekatan *One-Model-Per-Aspect* dengan Dataset *Balance*

Aspek	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Autentikasi Data	0.9361	0.9362	0.9342	0.9348
Verifikasi Data	0.9642	0.9649	0.9642	0.9640
Pengalaman Pengguna	0.9243	0.9246	0.9252	0.9249
Usability System	0.9143	0.9116	0.9116	0.9111
Layanan Pengguna	0.9586	0.9595	0.9577	0.9583
Rata-Rata	**0.9395**	**0.9393**	**0.9386**	**0.9386**

Variasi performa antar aspek tetap dipengaruhi oleh kekhasan istilah pada ulasan pengguna. Aspek Verifikasi Data mencatatkan kinerja tertinggi dengan *F1-Score* 96,42%. Keunggulan ini dipicu oleh penggunaan leksikon teknis yang sangat spesifik dan konsisten,

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15761

seperti *"kode EFIN"*, *"verifikasi email"*, atau *"kode OTP"*. Pola kata yang distingtif tersebut meminimalkan ambiguitas, sehingga model IndoBERT dapat mengidentifikasi konteks verifikasi secara akurat. Di sisi lain, aspek *Usability System* mencatatkan nilai terendah sebesar 91,11%, namun tetap berada dalam kategori performa yang sangat baik. Penurunan relatif ini disebabkan oleh variasi keluhan teknis yang cenderung generik dan sering kali tumpang tindih (*overlap*) dengan aspek lainnya.

Selain itu, reliabilitas hasil penelitian ini diperkuat melalui pengujian 10-*Fold Cross-Validation* pada aspek perwakilan (Autentikasi Data) sebagaimana disajikan pada Tabel 10. Hasil pengujian menunjukkan nilai standar deviasi yang sangat rendah, yaitu ±0,0032 pada dataset *balanced* dan ±0,0155 pada dataset *imbalanced* untuk metrik *Macro-F1*. Performa tertinggi pada kelas positif (*F1-score* 0,9669) menunjukkan bahwa arsitektur IndoBERT sangat efektif dalam mengenali pola linguistik ulasan kepuasan pengguna yang cenderung memiliki struktur kata yang lebih eksplisit dan seragam dibandingkan ulasan netral atau negatif. Rendahnya variansi ini membuktikan bahwa performa tinggi yang dicapai bersifat stabil dan konsisten terhadap perubahan distribusi data.

Pemilihan Aspek Autentikasi Data sebagai aspek representatif dalam pengujian stabilitas (10-*Fold Cross-Validation*) dan analisis kurva ROC didasarkan pada karakteristik distribusinya yang paling mencerminkan kondisi riil dataset. Aspek ini memiliki kompleksitas linguistik yang moderat—tidak terlalu teknis seperti aspek Verifikasi Data, namun tidak seambigu aspek *Usability System*—sehingga dianggap mampu memberikan gambaran performa model yang paling objektif terhadap keseluruhan data. Selain itu, penggunaan satu aspek representatif dilakukan sebagai strategi efisiensi komputasi tanpa menghilangkan validitas pengujian, mengingat pendekatan *One-Model-Per-Aspect* pada dasarnya menggunakan arsitektur dan konfigurasi *hyperparameter* yang identik untuk seluruh aspek lainnya.

Tabel 10. Detail Validasi Per-*Class* & Stabilitas (Aspek Autentikasi Data)

Kondisi	**Kelas**	***Precision***	***Recall***	***F1-Score***	***Macro-F1* (Mean ± STD)**
Imbalanced	Negatif	0.7712	0.7885	0.7797	0.8432 ± 0.0155
(Data Asli)	Netral	0.9205	0.9012	0.9107	
	Positif	0.8214	0.8448	0.8329	
Balanced	Negatif	0.9049	0.9455	0.9248	0.9359 ± 0.0032
(*Proposed*)	Netral	0.9555	0.8914	0.9223	
	Positif	0.95	0.9712	0.9605	

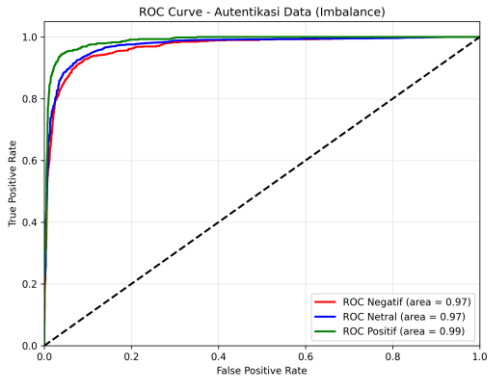
Analisis mendalam terhadap hasil klasifikasi menunjukkan bahwa pendekatan *One-Model-Per-Aspect* yang diusulkan memberikan peningkatan performa yang sangat signifikan dibandingkan dengan pendekatan *Single-Model-Multi-Aspect*. Berdasarkan data pada Tabel 11, terlihat adanya kenaikan rata-rata *F1-Score* sebesar 4,52%, dari 0,898 menjadi 0,9386. Peningkatan ini memvalidasi bahwa meskipun pendekatan model terpisah memerlukan biaya komputasi yang lebih tinggi, spesialisasi model terhadap masing-masing aspek mampu menangkap karakteristik linguistik yang unik dengan jauh lebih presisi.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15761

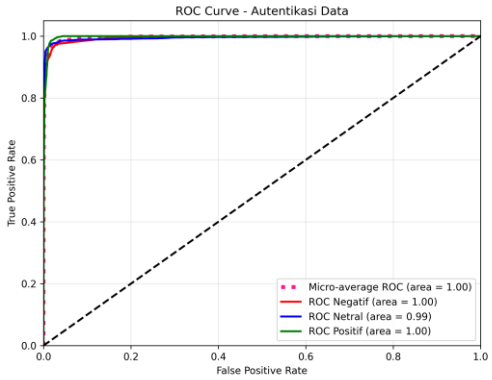
Tabel 11. Perbandingan Performa Metode pada Dataset *Balanced*

Pendekatan	***Avg. Accuracy***	***Avg. Precision***	***Avg. Recall***	***Avg. F1-Score***
Single-Model-Multi-Aspect	0.8986	0.8994	0.8985	0.8980
One-Model-Per-Aspect (*Proposed*)	0.9395	0.9393	0.9386	0.9386
Margin Peningkatan (%)	**+4.551**	**+4.436**	**+4.462**	**+4.521**

Secara keseluruhan, integrasi *fine-tuning* IndoBERT dengan strategi penyeimbangan data serta pendekatan *One-Model-Per-Aspect* terbukti menghasilkan sistem analisis sentimen yang jauh lebih *robust* dan akurat bagi aplikasi M-Pajak dibandingkan pendekatan model tunggal konvensional. Untuk memvalidasi ketegasan model dalam memisahkan antar kelas sentimen, dilakukan analisis visualisasi kurva ROC. Gambar 2 dan Gambar 3 menampilkan kurva ROC dari aspek representatif yakni Autentikasi Data dari penggunaan dataset seimbang maupun tidak seimbang.



Gambar 2. Kurva ROC-AUC Model IndoBERT Aspek Autentikasi Data pada Dataset *Imbalance*



Gambar 3. Kurva ROC-AUC Model IndoBERT Aspek Autentikasi Data pada Dataset *Balance*

Hasil menunjukkan bahwa semua kelas memiliki nilai AUC rata-rata (*micro-average*) di atas 99% pada dataset seimbang, yang menunjukkan bahwa model memiliki daya diskriminasi yang sangat tinggi (*high discriminatory power*). Pada Gambar 3, kurva menyentuh sudut kiri atas

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15761

secara baik (AUC 1,00), menandakan bahwa model paling andal dalam membedakan sentimen negatif dan positif dari sentimen lainnya tanpa kesalahan klasifikasi yang signifikan. Nilai sempurna ini dipicu oleh kemampuan IndoBERT dalam mengekstraksi fitur semantik yang sangat kontras setelah dataset diseimbangkan melalui *oversampling.* Sementara pada Gambar 2, meskipun terdapat sedikit variasi pada kelas netral dan negatif, model tetap mempertahankan nilai AUC diatas 0,95. Performa yang tinggi dan konsisten ini menegaskan bahwa model IndoBERT yang telah dioptimalkan melalui teknik *oversampling* sangat efektif untuk tugas klasifikasi sentimen multi-aspek pada ulasan aplikasi M-Pajak, bahkan pada data dengan karakteristik linguistik yang kompleks.

4. KESIMPULAN

Penelitian ini berhasil membuktikan efektivitas model *pre-trained* IndoBERT dalam melakukan analisis sentimen multi-aspek terhadap ulasan aplikasi layanan publik M-Pajak. Penentuan lima aspek utama dilakukan melalui metode *Latent Dirichlet Allocation* (LDA) yang divalidasi dengan nilai *Topic Coherence* optimal. Berdasarkan serangkaian eksperimen, konfigurasi *hyperparameter* optimal (*Epoch* 2, *Batch Size* 32, dan *Learning Rate* 5e-5) dengan pendekatan *One-Model-Per-Aspect* mampu menghasilkan rata-rata *Macro-F1* mencapai 93,86%. Reliabilitas model ini diperkuat dengan pengujian 10-*Fold Cross-Validation* yang menghasilkan standar deviasi sangat rendah (±0,0032), menunjukkan bahwa performa yang dicapai sangat stabil.

Temuan penting penelitian ini menyoroti peran krusial teknik *Random Oversampling* dalam menangani *class imbalance*. Tanpa penanganan ini, model cenderung bias terhadap kelas mayoritas (netral) dengan rata-rata *F1-score* hanya sebesar 84,07%. Analisis per aspek menunjukkan performa tertinggi pada aspek Verifikasi Data (*F1-score* 96,40%), sementara aspek *Usability System* (*F1-score* 91,11%) menjadi tantangan terbesar karena variasi gaya bahasa pengguna. Nilai AUC yang melebihi 0,99 pada dataset seimbang mengonfirmasi kemampuan arsitektur IndoBERT dalam memahami konteks semantik bahasa Indonesia yang kompleks serta efektivitas strategi penyeimbangan data dalam meminimalkan bias pada kelas minoritas.

DAFTAR PUSTAKA

[1] Direktorat Jenderal Pajak, “M-Pajak: Layanan pajak mudah dalam genggaman,” pajak.go.id.

[2] W. Astriningsih and D. Hatta Fudholi, “Identifikasi Multi Aspek Dan Sentimen Analisis Pada Review Hotel Menggunakan Deep Learning,” *Jatisi*, vol. 10, No. 3, September 2023, Hal. 433-443

[3] T. Hou, B. Yannou, Y. Leroy, and E. Poirson, “Mining customer product reviews for product development: A summarization process,” *Expert Syst. Appl.*, vol. 132, pp. 141–150, Oct. 2019, doi: 10.1016/j.eswa.2019.04.069.

[4] A. Maroof, S. Wasi, S. I. Jami, and M. S. Siddiqui, “Aspect-Based Sentiment Analysis for Service Industry,” *IEEE Access*, vol. 12, pp. 109702–109713, 2024, doi: 10.1109/ACCESS.2024.3440357.

[5] D. Wijaya, R. A. Saputra, and F. Irwiensyah, “Analisis Sentimen Ulasan Aplikasi Samsat Digital Nasional Pada Google Playstore Menggunakan Algoritma Naïve Bayes,” *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 4, no. 4, 2024, doi: 10.30865/klik.v4i4.1738.

[6] A. Faisol *et al.*, “Analisis Sentimen M-Pajak Menggunakan Algoritma Support Vector Machine, Naive Bayes dan KNN,” *Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)*, vol. 7, no. 5, 2024.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15761

[7] S. Jahan, M. Oussalah, D. Romaissa Beddia, J. Kabir Mim, and N. Arhab, "A Comprehensive Study on NLP Data Augmentation for Hate Speech Detection: Legacy Methods, BERT, and LLMs." 2024, doi: https://doi.org/10.48550/arXiv.2404.00303

[8] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," 2020, https://doi.org/10.48550/arXiv.2011.00677

[9] F. Ansyah and R. R. Suryono, "Sentiment Classification of Indonesian-Language Roblox Reviews Using IndoBERT with SMOTE Optimization", *JAIC*, vol. 9, no. 4, pp. 1868–1877, Aug. 2025.

[10] F. Y. A'la, "Optimasi Klasifikasi Sentimen Ulasan Game Berbahasa Indonesia: IndoBERT dan SMOTE untuk Menangani Ketidakseimbangan Kelas," *Edumatic: Jurnal Pendidikan Informatika*, vol. 9, no. 1, pp. 256–265, Apr. 2025, doi: 10.29408/edumatic.v9i1.29666.

[11] N. Alghamdi, S. Khatoon, and M. Alshamari, "Multi-Aspect Oriented Sentiment Classification: Prior Knowledge Topic Modelling and Ensemble Learning Classifier Approach," *Applied Sciences (Switzerland)*, vol. 12, no. 8, Apr. 2022, doi: 10.3390/app12084066.

[12] I. G. S. D. Putra and I. N. T. A. Putra, "IMPLEMENTASI METODE NAÏVE BAYES PADA ANALISIS SENTIMEN PENGGUNA APLIKASI MOBILE KITA BISA," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 2, Apr. 2025, doi: 10.23960/jitet.v13i2.6423.

[13] F. Agung, J. Ayomi, and K. E. Dewi, "ANALISIS EMOSI PADA MEDIA SOSIAL TWITTER MENGGUNAKAN METODE MULTINOMIAL NAÏVE BAYES DAN SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE," *KOMPUTA : Jurnal Ilmiah Komputer dan Informatika*, vol. 12, no. 2, 2023.

[14] D. Zakeshia Tiara Kannitha and P. Kartikasari, "PEMODELAN TOPIK PADA KELUHAN PELANGGAN MENGGUNAKAN ALGORITMA LATENT DIRICHLET ALLOCATION DALAM MEDIA SOSIAL TWITTER," vol. 11, no. 2, pp. 266–277, 2022.

[15] S. N. Almuayqil, M. Humayun, N. Z. Jhanjhi, M. F. Almufareh, and D. Javed, "Framework for Improved Sentiment Analysis via Random Minority Oversampling for User Tweet Review Classification," *Electronics (Switzerland)*, vol. 11, no. 19, Oct. 2022, doi: 10.3390/electronics11193058.

[16] L. Hakim, A. Sobri, L. Sunardi, and D. Nurdiansyah, "Prediksi penyakit jantung berbasis mesin learning dengan menggunakan metode k-nn," *Jurnal Digital Teknologi Informasi*, vol. 7, no. 2, p. 14, Feb. 2025, doi: 10.32502/digital.v7i2.9429.

[17] U. N. Khadijah and N. Cahyono, "Analisis Topic Modelling Pariwisata Yogyakarta Menggunakan Latent Dirichlet Allocation (LDA)," *The Indonesian Journal of Computer Science*, vol. 13, no. 4, pp. 6075-6086, Aug. 2024, doi: 10.33022/ijcs.v13i4.3816.

[18] M. Alexander and E. M. Sipayung, "User Reviews Sentiment Analysis on Internet Service Provider in Indonesia using Bidirectional Recurrent Neural Network," *8th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Aug. 2024, doi: 10.1109/ICITISEE63424.2024.10730617.

[19] A. P. Thenata and D. S. R. Saputra, "Analysis of Public Sentiment Towards President Prabowo's Work Program Using The CNN," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 4, pp. 1852–1857, Aug. 2025, doi: 10.30871/jaic.v9i4.9394.

[20] B. Hakim and P. R. Kinasih, "Sentiment Analysis of Indonesian Citizen Tweets Using Support Vector Machine on the Rebranding of Twitter to X," *Jurnal Tekno Kompak*, vol. 18, no. 2, p. 468, Jul. 2024, doi: 10.33365/jtk.v18i2.4293.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15761