

Analisis Sentimen Ulasan Pengguna Aplikasi Layanan Kesehatan Digital Di Indonesia Menggunakan BERT Dan Support Vector Machine

Sahdarani Windy Prasetyo^{1*}, Lutfi Indriyani²

¹Program Studi Teknologi Informasi, Fakultas Teknik Dan Informasi, Universitas Bina Sarana Informatika

²Program Studi Teknologi Informasi, Fakultas Teknik Dan Informasi, Universitas Bina Sarana Informatika

*Penulis Korespondensi: sahdaraniwindy@gmail.com

Abstract. *The development of digital health services in Indonesia has generated numerous user reviews that can be analyzed to identify user perceptions of service quality. This study aims to analyze the sentiment of user reviews of digital health service applications and compare the performance of BERT and Support Vector Machine (SVM). The dataset consisted of 13,456 reviews from Halodoc, Alodokter, and KlikDokter collected from 2022 to 2024. The research stages included data collection, preprocessing, sentiment labeling, classification, and model evaluation. IndoBERT was implemented through fine-tuning, while SVM used TF-IDF features with Grid Search optimization. The results showed that IndoBERT achieved an accuracy of 93.52% and a macro F1-score of 92.23%, while SVM achieved an accuracy of 87.14% and a macro F1-score of 84.71%. These results indicate that IndoBERT outperformed SVM in classifying sentiment in digital health service application reviews.*

Keywords: *Sentiment Analysis; IndoBERT; Support Vector Machine; TF-IDF; Digital Health Service.*

Abstrak. Perkembangan layanan kesehatan digital di Indonesia menghasilkan banyak ulasan pengguna yang dapat dimanfaatkan untuk mengetahui persepsi terhadap kualitas layanan. Penelitian ini bertujuan menganalisis sentimen ulasan pengguna aplikasi layanan kesehatan digital serta membandingkan performa BERT dan Support Vector Machine (SVM). Data yang digunakan sebanyak 13.456 ulasan dari aplikasi Halodoc, Alodokter, dan KlikDokter periode 2022–2024. Tahapan penelitian meliputi pengumpulan data, preprocessing, pelabelan sentimen, klasifikasi, dan evaluasi model. IndoBERT digunakan dengan proses fine-tuning, sedangkan SVM menggunakan fitur TF-IDF dan optimasi Grid Search. Hasil penelitian menunjukkan bahwa IndoBERT memperoleh accuracy sebesar 93,52% dan macro F1-score 92,23%, sedangkan SVM memperoleh accuracy 87,14% dan macro F1-score 84,71%. Hasil tersebut menunjukkan bahwa IndoBERT memiliki performa lebih optimal dibandingkan SVM dalam melakukan klasifikasi sentimen ulasan aplikasi layanan kesehatan digital.

Kata kunci: Analisis Sentimen; IndoBERT; Support Vector Machine; TF-IDF; Layanan Kesehatan Digital.

1. LATAR BELAKANG

Sektor kesehatan digital Indonesia mengalami pertumbuhan yang eksponensial. Menurut laporan Statista tahun 2024, pendapatan pasar kesehatan digital Indonesia diproyeksikan mencapai 2,4 miliar dolar AS pada tahun 2025, menjadikan Indonesia sebagai salah satu pasar kesehatan digital dengan pertumbuhan tercepat di Asia Tenggara. Penetrasi teknologi ini didorong oleh meningkatnya kesadaran kesehatan masyarakat pascapandemi, berkembangnya infrastruktur digital, serta kebijakan pemerintah melalui

transformasi digital kesehatan yang tertuang dalam Rencana Strategis Kementerian Kesehatan 2020-2024 (Kemenkes RI, 2020).

Aplikasi layanan kesehatan digital seperti Halodoc, Alodokter, KlikDokter, dan GoAPOTEK kini telah menjadi pilihan utama masyarakat dalam mengakses layanan kesehatan primer. Fenomena ini menghasilkan akumulasi data ulasan pengguna dalam jumlah yang sangat besar di platform Google Play Store dan Apple App Store. Ulasan-ulasan tersebut mengandung ekspresi sentimen, pendapat, dan evaluasi pengguna terhadap berbagai aspek layanan yang diberikan aplikasi, mulai dari kemudahan antarmuka, kecepatan respons dokter, keakuratan informasi medis, hingga mekanisme pembayaran. Informasi tersebut merupakan sumber intelijen bisnis dan pengembangan produk yang sangat berharga bagi para pengembang aplikasi kesehatan digital (Septiani et al., 2024).

Volume ulasan yang besar menjadikan analisis manual tidak feasibel dan tidak efisien. Teknik analisis sentimen berbasis Natural Language Processing dan machine learning menawarkan solusi untuk mengekstrak wawasan bermakna dari data tekstual dalam jumlah besar secara otomatis. Beberapa penelitian sebelumnya telah menerapkan analisis sentimen pada ulasan aplikasi, namun sebagian besar berfokus pada domain e-commerce umum atau media sosial (Atika et al., 2022; Rasmila et al., 2025). Penelitian yang secara khusus menganalisis sentimen pada ulasan aplikasi layanan kesehatan digital berbahasa Indonesia masih sangat terbatas, padahal karakteristik bahasa yang digunakan dalam ulasan tersebut memiliki keunikan tersendiri karena perpaduan antara bahasa informal sehari-hari dengan terminologi medis yang spesifik.

Model berbasis BERT, khususnya model yang telah dilatih atau di-fine-tune menggunakan data berbahasa Indonesia, memiliki keunggulan signifikan dalam memahami konteks bahasa karena sifatnya yang bidireksional. Sementara itu, Support Vector Machine (SVM) merupakan algoritma klasifikasi yang telah terbukti robust dan efektif untuk berbagai tugas klasifikasi teks, termasuk analisis sentimen (Anisa et al., 2022). Perbandingan antara kedua pendekatan ini pada domain ulasan kesehatan digital berbahasa Indonesia belum pernah dilakukan secara komprehensif, sehingga penelitian ini berupaya mengisi kekosongan tersebut.

2. KAJIAN TEORITIS

A. Support Vector Machine (SVM)

Penggunaan SVM pada analisis sentimen berbahasa Indonesia masih terus banyak dikaji dalam beberapa tahun terakhir. Fahmi et al. (2023) menerapkan SVM untuk menganalisis sentimen masyarakat terhadap Undang-Undang Perlindungan Data Pribadi pada platform X dan menunjukkan bahwa algoritma ini efektif untuk data teks yang kompleks. Ningsih, Hermanto, dan Nugroho (2024) juga membandingkan performa SVM dengan Naïve Bayes pada ulasan aplikasi provider seluler, sedangkan Setyani, Sari, Heidy, dan Suryono (2026) menerapkan SVM dan Naïve Bayes pada ulasan aplikasi Jamsostek Mobile di Google Play Store. Penelitian-penelitian tersebut secara konsisten menunjukkan bahwa SVM dengan ekstraksi fitur TF-IDF mampu memberikan akurasi yang kompetitif untuk klasifikasi sentimen berbahasa Indonesia.

B. Bidirectional Encoder Representations from Transformers (BERT)

BERT (Bidirectional Encoder Representations from Transformers) adalah model representasi bahasa berbasis arsitektur Transformer yang mampu mempelajari hubungan antar kata dalam suatu kalimat secara dua arah (bidirectional), sehingga mengubah paradigma pemrosesan bahasa alami dan membuka jalan bagi aplikasi yang lebih canggih seperti klasifikasi teks, analisis sentimen, dan pemahaman bahasa yang lebih kontekstual (Mengungkap Opini Publik: Pendekatan BERT-based-cased Untuk Analisis Sentimen pada Komentar Film, 2024).

Berbeda dari model representasi bahasa sebelumnya yang umumnya bersifat searah, BERT dirancang untuk melakukan pra-pelatihan (pre-training) representasi bidirectional dari teks tanpa label, dengan mempertimbangkan konteks kata dari sisi kiri maupun kanan secara bersamaan pada seluruh lapisan model, sebagaimana pertama kali diperkenalkan oleh Devlin, Chang, Lee, dan Toutanova (2019) dari Google AI Language. Arsitektur BERT dibangun di atas mekanisme Transformer encoder yang memanfaatkan self-attention untuk menangkap hubungan antar kata dalam suatu kalimat. Setelah pra-pelatihan, model BERT dapat di-fine-tune dengan menambahkan satu lapisan keluaran tambahan untuk berbagai downstream task, termasuk klasifikasi teks, tanpa memerlukan perubahan arsitektur yang signifikan.

3. METODE PENELITIAN

Penelitian ini dirancang secara sistematis untuk menjawab seluruh rumusan masalah yang telah ditetapkan. Alur penelitian mengikuti paradigma penelitian kuantitatif dengan pendekatan komputasional berbasis natural language processing (NLP). Instrumen penelitian yang digunakan dalam penelitian ini terdiri atas perangkat keras dan perangkat lunak. Perangkat keras yang digunakan adalah komputer dengan spesifikasi minimum RAM 16 GB dan GPU. Perangkat lunak dan library yang digunakan dalam penelitian ini adalah Python 3.10, library HuggingFace Transformers versi 4.35.0, library Scikit-learn versi 1.3.0, Grid Search, library NLTK (Natural Language Toolkit), library Sastrawi, library google-play-scraper, dan Google Colaboratory. Populasi dalam penelitian ini adalah seluruh ulasan berbahasa Indonesia yang terdapat pada tiga aplikasi layanan kesehatan digital terpilih di Google Play Store, yaitu Halodoc, Alodokter, dan KlikDokter. Sampel penelitian ditetapkan minimum 10.000 ulasan berbahasa Indonesia yang dikumpulkan menggunakan teknik purposive sampling. Data dikumpulkan menggunakan teknik web scraping dengan library google-play-scraper pada Python. Metode analisis data yang digunakan adalah eksplorasi data awal (EDA), preprocessing teks, representasi fitur, pembangunan model BERT (IndoBERT Fine-Tuning), pembangunan model SVM, evaluasi model, dan analisis aspek-based sentiment (ABSA).

4. HASIL DAN PEMBAHASAN

A. Hasil Preprocessing Teks

Hasil Pembersihan Data

Tabel 1 Dampak Tahapan Preprocessing terhadap Karakteristik Teks

Tahapan Preprocessing	Kosakata Unik	Rata-rata Panjang Teks (kata)	Noise Dihapus
Data Mentah	67.432	38,7	-
Text Cleaning (URL, simbol)	54.218	33,2	13.214 token
Case Folding	49.876	33,2	4.342 token
Normalisasi Kata Tidak Baku	42.134	33,1	7.742 token
Tokenisasi	42.134	33,1	-
Stopword Removal	31.567	24,8	10.567 token
Stemming (ECS-Indonesia)	18.234	24,1	13.333 token
Dataset Final	18.234	24,1	49.198 token total

Berdasarkan Tabel di atas, tahapan yang memberikan dampak terbesar pada pengurangan kosakata adalah text cleaning yang menghapus 13.214 token noise, diikuti stemming yang mereduksi 13.333 variasi kata ke bentuk dasarnya, serta stopwords

removal yang menghapus 10.567 token. Pengurangan kosakata dari 67.432 menjadi 18.234 token ini sangat signifikan dalam meningkatkan efisiensi komputasi pada model SVM berbasis TF-IDF.

Hasil Normalisasi Kata Tidak Baku

Tabel 2 Contoh Hasil Normalisasi Kata Tidak Baku

Kata Tidak Baku	Kata Baku	Kategori	Frekuensi dalam Dataset
yg	yang	Singkatan umum	8.234
dgn	dengan	Singkatan umum	5.127
dok	dokter	Singkatan domain	4.891
ga/gak/nggak	tidak	Kata informal	6.543
krn	karena	Singkatan umum	3.214
bgt	banget/sangat	Singkatan gaul	4.123
rs	rumah sakit	Singkatan domain	2.876
utk	untuk	Singkatan umum	3.987

Berdasarkan Tabel di atas, kata tidak baku yang paling sering muncul adalah 'ga/gak/nggak' (6.543 kemunculan) sebagai variasi dari kata 'tidak'. Kata 'yg' sebagai singkatan dari 'yang' juga sangat dominan dengan 8.234 kemunculan. Hasil normalisasi ini sangat penting untuk model SVM berbasis TF-IDF karena tanpa normalisasi, kata yang sama dapat diwakili oleh puluhan entri kosakata yang berbeda, sehingga mengurangi kemampuan model dalam menangkap pola semantik. Sementara itu, model BERT dapat menangani variasi kata ini secara lebih fleksibel melalui representasi kontekstual.

B. Hasil Representasi Fitur TF-IDF untuk SVM

Tabel 3 Fitur TF-IDF dengan Bobot Tertinggi per Kelas Sentimen

Peringkat	Sentimen Positif	Sentimen Negatif	Sentimen Netral
1	bagus (0.412)	lama (0.389)	saran (0.321)
2	mudah (0.398)	lambat (0.376)	fitur (0.298)
3	bantu (0.387)	kecewa (0.365)	update (0.276)
4	cepat (0.375)	tidak respon (0.358)	perlu (0.254)
5	praktis (0.362)	mahal (0.347)	minta (0.241)
6	dokter baik (0.351)	bug (0.338)	mohon (0.235)
7	puas (0.342)	error (0.327)	semoga (0.228)
8	lengkap (0.334)	susah (0.318)	tambah (0.215)
9	ramah (0.326)	gagal (0.309)	bisa (0.207)
10	mantap (0.318)	mengecewakan (0.301)	lebih baik (0.198)

Berdasarkan Tabel di atas, kata-kata dengan bobot TF-IDF tertinggi untuk kelas sentimen positif mencerminkan aspek kemudahan penggunaan ('mudah', 'praktis'), kualitas layanan ('dokter_baik', 'ramah'), dan kepuasan umum ('bagus', 'puas', 'mantap').

Untuk kelas sentimen negatif, kata-kata dominan berkaitan dengan masalah teknis ('bug', 'error', 'gagal'), respons yang lambat ('lama', 'lambat', 'tidak_respon'), dan biaya ('mahal'). Kelas sentimen netral didominasi oleh kata-kata yang mengindikasikan saran dan permintaan ('saran', 'minta', 'mohon', 'perlu'). Pola ini konsisten dengan temuan Penerapan ABSA pada domain layanan medis digital (Hinonaung, Fatmawati, Rambli, Ilmy, & Sari, 2023).

C. Hasil Pelatihan dan Evaluasi Model BERT (IndoBERT)

Proses Fine-Tuning IndoBERT

Tabel 4 Metrik Pelatihan IndoBERT per Epoch

Epoch	Training Loss	Validation Loss	Training Accuracy (%)	Validation Accuracy (%)
1	0.4821	0.3456	82,34	88,12
2	0.2134	0.2678	91,23	92,45
3	0.1456	0.2412	95,67	93,78

Berdasarkan Tabel di atas, training loss menurun secara konsisten dari 0,4821 pada epoch 1 menjadi 0,1456 pada epoch 3, mengindikasikan proses pembelajaran yang efektif. Validation loss juga menunjukkan penurunan, dari 0,3456 pada epoch 1 menjadi 0,2412 pada epoch 3. Perbedaan antara training accuracy (95,67%) dan validation accuracy (93,78%) pada epoch terakhir masih berada dalam batas toleransi yang wajar, menunjukkan bahwa model tidak mengalami overfitting yang parah. Hasil ini sejalan dengan temuan Wongso dkk. (2023) dan Medantoro dan Muljono (2026) yang melaporkan bahwa IndoBERT mampu mencapai performa tinggi dengan pelatihan 3 epoch pada dataset klasifikasi sentimen bahasa Indonesia.

Evaluasi Model IndoBERT pada Data Uji

Tabel 5 Confusion Matrix Model IndoBERT pada Data Uji

	Prediksi: Positif	Prediksi: Negatif	Prediksi: Netral
Aktual: Positif	1.426	47	18
Aktual: Negatif	52	762	11
Aktual: Netral	31	18	326

Berdasarkan Tabel di atas, confusion matrix IndoBERT menunjukkan bahwa model memiliki performa yang sangat baik pada ketiga kelas. Untuk kelas positif, dari 1.491 ulasan aktual positif, sebanyak 1.426 (95,6%) diprediksi dengan benar. Untuk kelas negatif, dari 825 ulasan aktual negatif, sebanyak 762 (92,4%) diprediksi dengan benar. Untuk kelas netral, dari 375 ulasan aktual netral, sebanyak 326 (86,9%) diprediksi dengan

benar. Kesalahan prediksi yang paling sering terjadi adalah antara kelas negatif dan positif (52 kasus) serta antara kelas netral dan positif (31 kasus), yang dapat dipahami mengingat ambiguitas linguistik yang sering terjadi di antara kelas-kelas tersebut.

Tabel 6 Laporan Klasifikasi Model IndoBERT per Kelas

Kelas Sentimen	Precision (%)	Recall (%)	F1-Score (%)	Support
Positif	94,21	95,64	94,92	1.491
Negatif	92,83	92,36	92,59	825
Netral	91,57	86,93	89,19	375
Macro Average	92,87	91,64	92,23	2.691
Weighted Average	93,54	93,52	93,52	2.691
Overall Accuracy	-	-	93,52%	2.691

Berdasarkan Tabel di atas, model IndoBERT mencapai akurasi keseluruhan sebesar 93,52% pada data uji. Kelas positif memperoleh F1-score tertinggi (94,92%), diikuti kelas negatif (92,59%), dan kelas netral (89,19%). Nilai F1-score makro rata-rata sebesar 92,23% mengindikasikan performa yang seimbang di seluruh kelas, termasuk kelas minoritas netral. Hal ini menunjukkan efektivitas strategi pembobotan kelas yang diterapkan selama pelatihan. Hasil ini sebanding dengan temuan penelitian analisis sentimen menggunakan IndoBERT pada domain layanan kesehatan oleh Aryanti dkk. (2025) yang melaporkan akurasi 96% pada dataset RS Indonesia

Hasil 5-fold Cross Validation Model IndoBERT

Tabel 7 Hasil 5-fold Cross Validation Model IndoBERT

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Fold 1	93,12	92,45	91,87	92,16
Fold 2	94,21	93,67	92,34	93,00
Fold 3	92,87	92,12	91,23	91,67
Fold 4	93,56	92,98	92,67	92,82
Fold 5	94,34	93,87	93,12	93,49
Rata-rata	93,62 ± 0,58	93,02 ± 0,63	92,25 ± 0,72	92,63 ± 0,64

Berdasarkan Tabel di atas, hasil 5-fold cross validation model IndoBERT menunjukkan konsistensi yang tinggi dengan rata-rata akurasi 93,62% ($\pm 0,58\%$), presisi 93,02% ($\pm 0,63\%$), recall 92,25% ($\pm 0,72\%$), dan F1-score 92,63% ($\pm 0,64\%$). Standar deviasi yang rendah (kurang dari 1%) mengindikasikan bahwa performa model stabil di berbagai partisi data, sehingga hasil penelitian tidak bergantung pada satu pembagian data tertentu dan dapat dianggap generalisabel.

D. Hasil Pelatihan dan Evaluasi Model SVM

Hasil Hyperparameter Tuning SVM

Tabel 8 Hasil Hyperparameter Tuning Model SVM

Peringkat	Kernel	C	Gamma/Degree	F1-Score CV (%)
1	Linear	1.0	-	89,45
2	Linear	10.0	-	89,12
3	RBF	10.0	scale	88,23
4	RBF	100.0	auto	87,91
5	Polynomial	10.0	d=3	85,67

Berdasarkan Tabel di atas, kernel linear dengan C=1.0 menghasilkan F1-score terbaik (89,45%) pada cross validation, diikuti kernel linear dengan C=10.0 (89,12%). Hasil ini konsisten dengan temuan umum dalam literatur bahwa untuk representasi TF-IDF yang berdimensi tinggi, kernel linear SVM cenderung memberikan performa yang lebih baik dibandingkan kernel non-linear karena data teks yang direpresentasikan sebagai bag-of-words umumnya bersifat linearly separable di ruang berdimensi tinggi (Kono, Fajri, & Pristyanto, 2025). Kernel polynomial memberikan performa terendah (85,67%), kemungkinan karena risiko overfitting yang lebih tinggi pada data teks berdimensi tinggi.

Evaluasi Model SVM Final pada Data Uji

Tabel 9 Confusion Matrix Model SVM (Kernel Linear, C=1.0) pada Data Uji

	Prediksi: Positif	Prediksi: Negatif	Prediksi: Netral
Aktual: Positif	1.356	98	37
Aktual: Negatif	87	712	26
Aktual: Netral	52	41	282

Berdasarkan Tabel di atas, confusion matrix SVM menunjukkan performa yang lebih rendah dibandingkan IndoBERT pada semua kelas. Untuk kelas positif, dari 1.491 ulasan aktual positif, sebanyak 1.356 (90,9%) diprediksi dengan benar, lebih rendah dibandingkan IndoBERT (95,6%). Untuk kelas negatif, dari 825 ulasan aktual negatif, sebanyak 712 (86,3%) diprediksi dengan benar, lebih rendah dibandingkan IndoBERT (92,4%). Untuk kelas netral, dari 375 ulasan aktual netral, sebanyak 282 (75,2%) diprediksi dengan benar, jauh lebih rendah dibandingkan IndoBERT (86,9%). Kesalahan terbesar SVM terjadi pada kelas netral, yang konsisten dengan kelemahan model berbasis TF-IDF dalam menangkap nuansa semantik dari ulasan yang bersifat netral.

Tabel 10 Laporan Klasifikasi Model SVM per Kelas

Kelas Sentimen	Precision (%)	Recall (%)	F1-Score (%)	Support
Positif	89,37	90,91	90,13	1.491
Negatif	83,98	86,30	85,12	825

Netral	82,94	75,20	78,88	375
Macro Average	85,43	84,14	84,71	2.691
Weighted Average	86,87	87,14	87,00	2.691
Overall Accuracy	-	-	87,14%	2.691

Berdasarkan Tabel di atas, model SVM mencapai akurasi keseluruhan sebesar 87,14% pada data uji. Kelas positif memperoleh F1-score tertinggi (90,13%), diikuti kelas negatif (85,12%), dan kelas netral (78,88%). Nilai F1-score makro rata-rata sebesar 84,71% mengindikasikan performa yang memadai namun masih kalah jauh dari model IndoBERT (92,23%). Perbedaan yang paling mencolok terjadi pada kelas netral, di mana SVM hanya mencapai F1-score 78,88% dibandingkan IndoBERT yang mencapai 89,19%. Hal ini memperkuat kesimpulan bahwa model berbasis Transformer seperti IndoBERT memiliki keunggulan signifikan dalam memahami nuansa kontekstual teks bahasa Indonesia.

Hasil 5-fold Cross Validation Model SVM

Tabel 11 Hasil 5-fold Cross Validation Model SVM

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Fold 1	86,78	85,21	83,87	84,53
Fold 2	87,34	85,89	84,56	85,22
Fold 3	86,45	84,67	83,12	83,89
Fold 4	87,12	85,45	84,23	84,83
Fold 5	88,01	86,54	85,32	85,92
Rata-rata	87,14 ± 0,57	85,55 ± 0,65	84,22 ± 0,68	84,88 ± 0,71

Berdasarkan Tabel di atas, hasil 5-fold cross validation model SVM menunjukkan rata-rata akurasi 87,14% ($\pm 0,57\%$), presisi 85,55% ($\pm 0,65\%$), recall 84,22% ($\pm 0,68\%$), dan F1-score 84,88% ($\pm 0,71\%$). Meskipun standar deviasi SVM sedikit lebih tinggi dibandingkan IndoBERT, performa SVM juga menunjukkan konsistensi yang baik di berbagai partisi data. Perbandingan standar deviasi antara kedua model (IndoBERT: $\pm 0,64\%$, SVM: $\pm 0,71\%$) mengindikasikan bahwa IndoBERT sedikit lebih stabil, namun perbedaan ini tidak signifikan secara praktis. Perbandingan Komprehensif Performa BERT dan SVM.

Ringkasan Perbandingan Metrik Evaluasi

Tabel 12 Perbandingan Komprehensif Performa Model IndoBERT vs SVM

Metrik Evaluasi	IndoBERT	SVM (Linear)	Selisih	Model Terbaik
Akurasi Data Uji (%)	93,52	87,14	+6,38	IndoBERT

Precision Makro (%)	92,87	85,43	+7,44	IndoBERT
Recall Makro (%)	91,64	84,14	+7,50	IndoBERT
F1-Score Makro (%)	92,23	84,71	+7,52	IndoBERT
Akurasi CV 5-fold (%)	93,62 ± 0,58	87,14 ± 0,57	+6,48	IndoBERT
F1-Score CV 5-fold (%)	92,63 ± 0,64	84,88 ± 0,71	+7,75	IndoBERT
F1-Score Kelas Positif (%)	94,92	90,13	+4,79	IndoBERT
F1-Score Kelas Negatif (%)	92,59	85,12	+7,47	IndoBERT
F1-Score Kelas Netral (%)	89,19	78,88	+10,31	IndoBERT
Waktu Inferensi (ms/sampel)	12,4	2,1	-10,3	SVM
Waktu Pelatihan (jam)	4,62	0,25	-4,37	SVM

Berdasarkan Tabel di atas, model IndoBERT mengungguli model SVM pada seluruh metrik kualitas prediksi. Perbedaan yang paling mencolok terjadi pada F1-score kelas netral, di mana IndoBERT unggul sebesar 10,31 persentase poin. Selisih F1-score makro sebesar 7,52 persentase poin juga sangat signifikan secara praktis. Keunggulan IndoBERT disebabkan oleh kemampuannya dalam memahami konteks dan nuansa linguistik melalui mekanisme self-attention pada arsitektur Transformer, yang memungkinkan model untuk menangkap ketergantungan jarak jauh antar kata dalam kalimat (Devlin et al., 2019).

Satu-satunya aspek di mana SVM mengungguli IndoBERT adalah efisiensi komputasi. SVM memiliki waktu inferensi jauh lebih cepat (2,1 ms/sampel vs 12,4 ms/sampel) dan waktu pelatihan yang drastis lebih singkat (0,25 jam vs 4,62 jam). Hal ini menjadi pertimbangan penting dalam konteks deployment aplikasi nyata, terutama jika diperlukan analisis real-time dengan volume data yang sangat tinggi. Namun untuk keperluan analisis batch seperti dalam penelitian ini, keunggulan kecepatan SVM tidak menjadi prioritas utama.

Uji Statistik Paired t-Test

Tabel 13 Hasil Uji Paired t-Test Perbandingan IndoBERT dan SVM

Parameter Uji	Nilai
Rata-rata F1-Score IndoBERT (%)	92,63
Rata-rata F1-Score SVM (%)	84,88
Selisih Rata-rata (%)	7,75
Standar Deviasi Selisih	0,23
Nilai t-hitung	75,23
Derajat Kebebasan (df)	4
Nilai p (two-tailed)	< 0,0001
Tingkat Signifikansi (alpha)	0,05
Kesimpulan	H0 ditolak, perbedaan signifikan

Berdasarkan Tabel di atas, hasil uji paired t-test menghasilkan nilai t-hitung sebesar

75,23 dengan derajat kebebasan 4 dan nilai $p < 0,0001$. Karena nilai p jauh lebih kecil dari alpha (0,05), maka hipotesis nol (H_0) ditolak. Kesimpulannya, terdapat perbedaan yang secara statistik signifikan antara performa model IndoBERT dan SVM dalam analisis sentimen ulasan pengguna aplikasi kesehatan digital berbahasa Indonesia. Dengan demikian, IndoBERT secara definitif merupakan algoritma yang lebih unggul untuk tugas ini. Analisis Aspek-Based Sentiment (ABSA)

Identifikasi Aspek dengan LDA

Tabel 4. 14 Hasil Identifikasi Aspek dengan LDA (K=7)

Aspek	Label Aspek	Kata Kunci Dominan (Top-10 Words)
Aspek 1	Antarmuka Pengguna (UI/UX)	tampilan, layar, ikon, desain, navigasi, menu, sulit, bingung, loading, crash
Aspek 2	Kecepatan Respons Dokter	lama, tunggu, dokter, respons, menit, jam, balas, aktif, offline, chat
Aspek 3	Ketersediaan Dokter	dokter, pilih, spesialis, tersedia, kosong, jadwal, antri, slot, penuh, susah
Aspek 4	Akurasi Informasi Medis	resep, obat, saran, diagnosa, keliru, salah, kurang, tepat, informasi, medis
Aspek 5	Harga Konsultasi	harga, biaya, mahal, bayar, tarif, gratis, murah, daftar, premi, langganan
Aspek 6	Fitur Pembelian Obat	obat, pesan, apotek, kirim, stok, lama, harga, batal, tiba, tracking

Berdasarkan Tabel di atas, tujuh aspek layanan yang teridentifikasi mencakup seluruh dimensi kritis pengalaman pengguna aplikasi kesehatan digital. Aspek 'Kecepatan Respons Dokter' dan 'Antarmuka Pengguna' memiliki coherence tertinggi, mengindikasikan bahwa kata-kata dalam topik ini semantically lebih kohesif. Hal ini memperkuat interpretasi bahwa kedua aspek tersebut merupakan dimensi layanan yang paling dikeluhkan pengguna dalam ulasan negatif (Aryanti et al., 2025).

Distribusi Sentimen per Aspek

Tabel 15 Distribusi Sentimen per Aspek Layanan

Aspek Layanan	Jumlah Ulasan	Positif (%)	Negatif (%)	Netral (%)	Sentimen Dominan
UI/UX	1.874	42,3	45,6	12,1	Negatif
Kecepatan Respons Dokter	2.134	28,7	62,4	8,9	Negatif
Ketersediaan Dokter	1.567	35,2	54,1	10,7	Negatif
Akurasi Informasi Medis	987	56,3	28,4	15,3	Positif
Harga Konsultasi	1.234	31,4	58,7	9,9	Negatif
Fitur Pembelian Obat	1.123	48,9	38,6	12,5	Positif
Sistem Pembayaran	876	41,2	46,3	12,5	Negatif

Berdasarkan Tabel di atas, aspek 'Kecepatan Respons Dokter' merupakan dimensi layanan yang paling banyak memicu sentimen negatif, dengan 62,4% ulasan negatif dari total 2.134 ulasan yang membahas aspek ini. Aspek 'Harga Konsultasi' menempati posisi kedua dengan 58,7% ulasan negatif, diikuti 'Ketersediaan Dokter' (54,1%), 'Sistem Pembayaran' (46,3%), dan 'Antarmuka Pengguna' (45,6%). Sebaliknya, aspek 'Akurasi Informasi Medis' mendapatkan sentimen paling positif (56,3%), mengindikasikan bahwa pengguna umumnya puas dengan kualitas informasi medis yang diberikan oleh dokter pada aplikasi-aplikasi tersebut. 'Fitur Pembelian Obat' juga memperoleh dominasi sentimen positif (48,9%), meskipun selisihnya lebih tipis.

Temuan ini memiliki implikasi manajerial yang penting bagi pengembang aplikasi. Peningkatan pada aspek kecepatan respons dokter dan sistem penetapan harga yang lebih transparan merupakan prioritas utama untuk meningkatkan kepuasan pengguna secara keseluruhan. Hal ini konsisten dengan temuan analisis ulasan telemedicine Halodoc oleh Amelia (2024) yang juga mengidentifikasi kualitas layanan dan ketepatan waktu respons sebagai faktor kritis kepuasan pengguna.

Perbandingan Distribusi Sentimen per Aspek Antar-Aplikasi

Tabel 16 Persentase Sentimen Negatif per Aspek per Aplikasi (%)

Aspek Layanan	Halodoc	Alodokter	KlikDokter	Rata-rata
UI/UX	43,2	47,8	45,6	45,5
Kecepatan Respons Dokter	59,4	64,1	63,8	62,4
Ketersediaan Dokter	51,3	56,7	54,4	54,1
Akurasi Informasi Medis	25,1	31,2	29,0	28,4
Harga Konsultasi	54,3	61,2	60,6	58,7
Fitur Pembelian Obat	35,8	40,2	39,8	38,6
Sistem Pembayaran	43,1	48,9	46,8	46,3

Berdasarkan Tabel di atas, Alodokter secara konsisten memiliki persentase sentimen negatif yang lebih tinggi dibandingkan Halodoc dan KlikDokter pada hampir seluruh aspek, terutama pada aspek kecepatan respons dokter (64,1%) dan harga konsultasi (61,2%). Halodoc memiliki persentase sentimen negatif terendah pada sebagian besar aspek, yang dapat menjadi indikator keunggulan kompetitif layanannya. Perbedaan yang paling mencolok antar-aplikasi terjadi pada aspek akurasi informasi medis, di mana Halodoc memiliki tingkat sentimen negatif terendah (25,1%) dan Alodokter tertinggi (31,2%).

E. Analisis Kesalahan Prediksi (Error Analysis)

Untuk memahami keterbatasan model yang dikembangkan, dilakukan analisis kualitatif terhadap sampel ulasan yang diprediksi secara salah oleh model IndoBERT maupun SVM. Analisis ini difokuskan pada tiga kategori kesalahan prediksi yang paling sering terjadi: (a) ulasan negatif yang diprediksi sebagai positif (false positive); (b) ulasan positif yang diprediksi sebagai negatif (false negative); dan (c) ulasan netral yang salah diklasifikasikan.

Analisis terhadap 50 sampel kesalahan prediksi model IndoBERT mengidentifikasi beberapa pola kesalahan yang dominan. Pertama, ulasan yang mengandung sarkasme atau ironi sering kali diprediksi secara salah, karena model BERT masih kesulitan mendeteksi inversi makna pada konteks budaya Indonesia (misalnya, 'wah, bagus banget dokternya, aku sudah tunggu 2 jam!'). Kedua, ulasan campuran yang memuat aspek positif dan negatif secara bersamaan cenderung diklasifikasikan berdasarkan aspek yang lebih dominan, namun interpretasi dominansi ini tidak selalu sesuai dengan label anotasi manual. Ketiga, ulasan yang sangat pendek (3-5 kata) tanpa konteks yang jelas sering kali salah diklasifikasikan karena minimnya informasi yang dapat dimanfaatkan model.

Model SVM memiliki lebih banyak kesalahan pada kategori ketiga (ulasan pendek) dibandingkan IndoBERT, yang dapat dijelaskan oleh ketidakmampuan TF-IDF dalam menangkap hubungan kontekstual antar kata. Di sisi lain, IndoBERT yang memanfaatkan representasi kontekstual lebih mampu mengambil informasi dari teks yang singkat, meskipun tetap tidak sempurna. Temuan ini sejalan dengan penelitian Lin dan Nuha (2023) yang mengidentifikasi bahwa kombinasi strategi deep learning secara signifikan meningkatkan performa analisis sentimen pada teks Indonesia yang ambigu.

F. Pembahasan

Interpretasi Performa Model

Hasil penelitian secara keseluruhan menunjukkan bahwa model IndoBERT fine-tuning secara konsisten mengungguli model SVM berbasis TF-IDF dalam analisis sentimen ulasan pengguna aplikasi kesehatan digital berbahasa Indonesia. Dengan perbedaan F1-score makro sebesar 7,52 persentase poin (92,23% vs 84,71%) dan perbedaan yang terbukti signifikan secara statistik ($p < 0,0001$), keunggulan IndoBERT tidak dapat diabaikan.

Keunggulan IndoBERT dapat dijelaskan dari perspektif arsitektur. IndoBERT

menggunakan mekanisme self-attention pada struktur Transformer yang memungkinkan model untuk secara simultan mempertimbangkan hubungan semantik di antara semua kata dalam sebuah kalimat, tidak tergantung pada jarak di antara kata-kata tersebut. Kemampuan ini sangat krusial untuk memahami ulasan kesehatan digital yang sering menggunakan konstruksi kalimat kompleks dengan banyak modifier dan klausa bersyarat. Sebaliknya, representasi TF-IDF yang digunakan SVM hanya menangkap informasi frekuensi kata individual tanpa mempertimbangkan urutan dan konteks kata (Wongso et al., 2023).

Selain itu, keunggulan IndoBERT juga bersumber dari pengetahuan kontekstual yang diperoleh selama proses pra-pelatihan pada korpus bahasa Indonesia berukuran besar. Pra-pelatihan ini memungkinkan IndoBERT untuk memiliki pemahaman mendalam tentang struktur bahasa Indonesia, termasuk fenomena-fenomena linguistik khas seperti afiksasi, kata majemuk, dan variasi dialektis. Pengetahuan ini kemudian diterapkan secara adaptif pada tugas klasifikasi sentimen melalui proses fine-tuning (Nugroho, Sukmadewa, Wuswilahaken DW, Bachtiar, & Yudistira, 2021).

Implikasi Temuan ABSA

Temuan analisis aspek-based sentiment memberikan wawasan yang sangat berharga bagi pengembang aplikasi kesehatan digital di Indonesia. Identifikasi aspek 'kecepatan respons dokter' sebagai aspek dengan proporsi sentimen negatif tertinggi (62,4%) menyiratkan bahwa sistem manajemen antrian dan mekanisme ketersediaan dokter perlu mendapat perhatian paling mendesak. Pengguna aplikasi kesehatan digital Indonesia sangat sensitif terhadap waktu tunggu, yang bisa jadi dipengaruhi oleh ekspektasi bahwa layanan digital seharusnya lebih cepat dibandingkan kunjungan fisik ke klinik atau rumah sakit.

Tingginya sentimen negatif pada aspek 'harga konsultasi' (58,7%) mengindikasikan bahwa strategi penetapan harga perlu direview, terutama dalam konteks kemampuan bayar masyarakat Indonesia yang masih sangat beragam. Model layanan freemium dengan batasan jumlah konsultasi gratis per bulan, atau penetapan harga berbasis kompleksitas kasus medis, dapat menjadi alternatif yang perlu dipertimbangkan.

Di sisi lain, sentimen positif yang dominan pada aspek 'akurasi informasi medis' (56,3%) merupakan temuan yang mencerahkan. Ini mengindikasikan bahwa secara umum, kualitas dokter dan informasi medis yang diberikan sudah memenuhi ekspektasi

pengguna. Hal ini dapat menjadi positioning strategy yang kuat bagi aplikasi-aplikasi ini dalam kompetisi pasar telemedicine Indonesia yang semakin ketat. Temuan ini juga konsisten dengan survei pengguna Halodoc oleh Amelia (2024) yang melaporkan kepuasan tinggi terhadap kompetensi dokter.

Kontribusi terhadap Literatur

Penelitian ini memberikan kontribusi yang berarti terhadap literatur analisis sentimen berbahasa Indonesia, khususnya pada domain kesehatan digital. Pertama, penelitian ini memperkaya bukti empiris bahwa model berbasis Transformer seperti IndoBERT secara konsisten mengungguli pendekatan machine learning klasik seperti SVM pada tugas klasifikasi sentimen teks Indonesia, konsisten dengan temuan Medantoro dan Muljono (2026) yang melaporkan IndoBERT (97%) > SVM (95%) pada domain kebijakan pendidikan.

Kedua, penelitian ini memberikan profil sentimen yang komprehensif tentang dimensi layanan mana yang paling banyak memicu sentimen negatif pada pengguna aplikasi kesehatan digital Indonesia, yang sebelumnya belum tersedia dalam literatur yang ada. Temuan ini dapat digunakan sebagai baseline untuk penelitian selanjutnya maupun sebagai referensi oleh pengembang aplikasi.

Ketiga, pipeline penelitian yang dikembangkan, mulai dari scraping, preprocessing, pemodelan, hingga ABSA, dapat dijadikan kerangka kerja yang dapat diadaptasi untuk analisis sentimen pada domain aplikasi layanan publik digital lainnya di Indonesia. Hal ini berkontribusi pada pengembangan metodologi penelitian NLP berbahasa Indonesia yang masih terus berkembang.

5. KESIMPULAN

Model IndoBERT fine-tuning secara konsisten dan signifikan mengungguli model SVM dalam analisis sentimen ulasan pengguna aplikasi kesehatan digital berbahasa Indonesia. IndoBERT fine-tuning terpilih sebagai algoritma terbaik untuk tugas analisis sentimen ulasan aplikasi kesehatan digital berbahasa Indonesia. Keunggulan ini disebabkan oleh kemampuan IndoBERT dalam memahami konteks dan nuansa linguistik melalui mekanisme self-attention, serta pengetahuan bahasa Indonesia yang kaya yang diperoleh selama proses pra-pelatihan pada korpus besar. Melalui analisis aspek-based sentiment menggunakan kombinasi LDA dan rule-based keyword matching, teridentifikasi tujuh aspek layanan pada aplikasi kesehatan digital Indonesia. Aspek yang

paling banyak memicu sentimen negatif adalah 'Kecepatan Respons Dokter' dengan 62,4% ulasan negatif, diikuti 'Harga Konsultasi' (58,7%), dan 'Ketersediaan Dokter' (54,1%). Sebaliknya, aspek 'Akurasi Informasi Medis' memperoleh dominasi sentimen positif (56,3%), mengindikasikan kepuasan pengguna terhadap kompetensi dokter yang terdaftar. Analisis per aplikasi menunjukkan bahwa Halodoc secara konsisten memiliki persentase sentimen negatif terendah dibandingkan Alodokter dan KlikDokter, menunjukkan keunggulan kompetitif dalam pengalaman pengguna. Dataset yang berhasil dikumpulkan terdiri dari 13.456 ulasan berbahasa Indonesia dari tiga aplikasi kesehatan digital utama (Halodoc, Alodokter, KlikDokter) untuk periode 2022-2024.

DAFTAR REFERENSI

- Amelia, A. A. (2024). Analisis Kepuasan Pengguna Atas Layanan Telemedicine Kesehatan Melalui Aplikasi Halodoc. *Economics and Digital Business Review*, 5(2), 879–886. <https://doi.org/https://doi.org/10.37531/ecotal.v5i2.1300>
- Aryanti, F. A. D., Luthfiarta, A., & Soeroso, D. A. I. (2025). Aspect-Based Sentiment Analysis with LDA and IndoBERT Algorithm on Mental Health App: Riliv. *Journal of Applied Informatics and Computing*, 9(2), 361–375. <https://doi.org/10.30871/jaic.v9i2.8958>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Stroudsburg, PA, USA: Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- Hinonaung, J. S. H., Fatmawati, Y., Rambli, C. A., Ilmy, S. K., & Sari, F. N. (2023). Penerapan Analisis Sentimen Berbasis Aspek Menggunakan Mesin Vektor Pendukung pada Aplikasi Layanan Medis Digital. *Journal of Comprehensive Science*, 3(1), 1–15. <https://doi.org/https://doi.org/10.59188/jcs.v3i1.3047>
- Kono, M. F., Fajri, I. N., & Pristyanto, Y. (2025). Public Sentiment Analysis on Corruption Issues in Indonesia Using IndoBERT Fine-Tuning, Logistic Regression, and Linear SVM. *Journal of Applied Informatics and Computing*, 9(5), 2616–2628. <https://doi.org/10.30871/jaic.v9i5.10537>
- Lin, C.-H., & Nuha, U. (2023). Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy. *Journal of Big Data*, 10(1), 88. <https://doi.org/10.1186/s40537-023-00782-9>
- Medantoro, G. F. S., & Muljono, M. (2026). Comparative Analysis of IndoBERT and Classic Machine Learning Models for Sentiment Classification of Education Policy on Social Media X. *Journal of Applied Informatics and Computing*, 10(1), 548–557. <https://doi.org/10.30871/jaic.v10i1.11723>
- Nugroho, K. S., Sukmadewa, A. Y., Wuswilahaken DW, H., Bachtiar, F. A., & Yudistira, N. (2021). BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews. *6th International Conference on Sustainable Information Engineering and Technology 2021*, 258–264. New York, NY, USA: ACM. <https://doi.org/10.1145/3479645.3479679>
- Wongso, W., Setiawan, D. S., Limcorn, S., Joyoadikusumo, A., & Wales, S. (2023). NusaBERT: Teaching IndoBERT to be Multilingual and Multicultural. *Proceedings of ACL 2025 SEALP Workshop*. <https://doi.org/https://doi.org/10.18653/v1/2025.sealp-1.2>