

Perbandingan Interpretabilitas Model Klasik *Machine Learning* dan *IndoBERT* pada Analisis Sentimen Aplikasi Layanan Kesehatan di Indonesia menggunakan Pengujian Generalisasi dengan Multi-Ulasan

Comparing the Interpretability of Classical Machine Learning Models and IndoBERT in Sentiment Analysis of Indonesian Health Service Applications using Multi-Review Generalization Testing

¹Muhammad Ridlo Nu'man Hakim*, ²Dwi Hosanna Bangkalang

^{1,2}Program Studi Sistem Informasi, Fakultas Teknologi Informasi, Universitas Kristen Satya Wacana

^{1,2}Jl. Dr. O. Notohamidjojo Blotongan, Sidorejo, Kota Salatiga, Jawa Tengah, Indonesia

*e-mail: 682022109@student.uksw.edu

(received: 26 August 2026, revised: 28 August 2026, accepted: 29 August 2026)

Abstrak

Meningkatnya penggunaan aplikasi layanan kesehatan di Indonesia turut mendorong pertumbuhan ulasan pengguna di Google Play Store. Kondisi ini menjadikan analisis sentimen secara manual semakin tidak efisien, terlebih pada ulasan yang memiliki bahasa beragam dan kalimat kontradiktif yang mengandung pergeseran sentimen antarklausa dan rentan menyebabkan kesalahan klasifikasi. Namun, kemampuan model dalam menangani kalimat kontradiktif tersebut masih jarang dikaji secara khusus. Di sisi lain, penelitian analisis sentimen terdahulu cenderung hanya melaporkan metrik performa tanpa turut mengevaluasi kemampuan generalisasi dan interpretabilitas dari sebuah model. Penelitian ini mengevaluasi model *machine learning* klasik *Naïve Bayes*, *Logistic Regression*, *SVM*, *Random Forest*, *XGBoost* dan model *transformer* *IndoBERT* dalam klasifikasi sentimen ulasan aplikasi layanan kesehatan berbahasa Indonesia, khususnya pada kalimat kontradiktif melalui perbandingan performa, pengujian generalisasi *out-of-distribution*, serta analisis *Integrated Gradients*. Metode penelitian mencakup pengumpulan 20.000 ulasan dari Google Play Store, pelabelan yang divalidasi terhadap pelabelan manual, pemodelan menggunakan model klasik dan *fine-tuning* untuk *IndoBERT*, serta pengujian generalisasi *out-of-distribution*. Hasil penelitian menunjukkan bahwa *IndoBERT* memperoleh F1-score tertinggi sebesar 98,10%, lebih tinggi dibandingkan model klasik sebesar 96,28%–97,57%. Pada *contrastive review*, *IndoBERT* menunjukkan performa generalisasi yang konsisten dengan akurasi 92,50% pada skenario *in-domain* dan 77,50% pada *cross-domain*, dibandingkan model klasik yang direntang 65,00%–72,50% dan 67,50%–75,00%. Analisis lebih lanjut menggunakan *Integrated Gradients* menunjukkan bahwa keputusan *IndoBERT* lebih dipengaruhi oleh token pembawa sentimen utama, terutama kata sifat dan negasi dibandingkan konjungsi kontradiktif itu sendiri. Temuan ini dapat menjadi referensi untuk pengembangan model analisis sentimen berbahasa Indonesia serta mendukung pengembang aplikasi layanan kesehatan dalam mengevaluasi kualitas layanan berdasarkan persepsi pengguna.

Kata kunci: analisis sentimen, *IndoBERT*, *integrated gradients*, interpretabilitas, pengujian generalisasi

Abstract

The increasing use of healthcare service applications in Indonesia has led to a growing volume of user reviews on the Google Play Store. This development makes manual sentiment analysis increasingly inefficient, particularly for reviews containing diverse linguistic expressions and contrastive sentences that involve sentiment shifts across clauses and are prone to causing classification errors. However, the ability of sentiment analysis models to handle such contrastive sentences has received limited attention. Moreover, previous sentiment analysis studies have tended to report performance metrics without systematically evaluating model generalization and interpretability. This study evaluates classical machine learning models, namely *Naïve Bayes*,

<http://sistemasi.ftik.unisi.ac.id>

Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost, alongside the IndoBERT transformer model for sentiment classification of Indonesian-language reviews of healthcare service applications, with particular emphasis on contrastive sentences. The evaluation covers performance comparison, out-of-distribution generalization testing, and interpretability analysis using Integrated Gradients. The research methodology includes collecting 20,000 reviews from the Google Play Store, labeling the reviews with validation against manually assigned labels, developing models using classical machine learning algorithms and fine-tuning IndoBERT, and conducting out-of-distribution generalization testing. The results show that IndoBERT achieved the highest F1-score of 98.10%, outperforming the classical machine learning models, which achieved F1-scores ranging from 96.28% to 97.57%. On contrastive reviews, IndoBERT demonstrated more consistent generalization performance, achieving an accuracy of 92.50% in the in-domain scenario and 77.50% in the cross-domain scenario, compared with ranges of 65.00%–72.50% and 67.50%–75.00%, respectively, for the classical models. Further analysis using Integrated Gradients revealed that IndoBERT's predictions were influenced more strongly by sentiment-bearing tokens, particularly adjectives and negations, than by contrastive conjunctions themselves. These findings provide a reference for developing Indonesian-language sentiment analysis models and support healthcare service application developers in evaluating service quality based on user perceptions.

Keywords: *generalization test, IndoBERT, integrated gradients, interpretability, sentiment analysis*

1 Pendahuluan

Kemajuan teknologi telah mendorong transformasi di berbagai sektor kehidupan, salah satunya pada sektor kesehatan. Pemanfaatan teknologi informasi dan komunikasi telah mempercepat digitalisasi layanan kesehatan melalui penerapan berbagai layanan berbasis digital untuk meningkatkan efisiensi operasional, akurasi pengelolaan data, serta kualitas pelayanan kepada masyarakat. Implementasi tersebut dapat diwujudkan melalui penerapan rekam medis elektronik, layanan *telemedicine*, dan integrasi sistem informasi kesehatan [1]. Salah satu bentuk implementasinya adalah berkembangnya aplikasi kesehatan berbasis *mobile* yang memungkinkan masyarakat memperoleh akses terhadap berbagai layanan kesehatan dengan lebih mudah, cepat, dan efisien [2]. Di Indonesia, transformasi tersebut diwujudkan melalui pengembangan berbagai aplikasi kesehatan oleh pemerintah maupun sektor swasta. Pemerintah menghadirkan Mobile JKN sebagai aplikasi resmi bagian dari program Jaminan Kesehatan Nasional yang menyediakan berbagai layanan, seperti pengelolaan data administrasi kesehatan, akses layanan medis, serta layanan informasi penting terkait program jaminan kesehatan [3]. Selain itu, pemerintah juga mengembangkan SatuSehat Mobile sebagai pengembangan dari aplikasi PeduliLindungi yang kini berfungsi menjadi *platform* layanan kesehatan terintegrasi dengan fitur rekam medis elektronik, hasil pemeriksaan tes laboratorium kesehatan, informasi imunisasi anak, serta akses informasi tentang fasilitas pelayanan kesehatan [4]. Di sisi lain, sektor swasta menghadirkan Halodoc sebagai aplikasi *telemedicine* yang menyediakan layanan konsultasi dokter, pembelian obat, serta berbagai layanan kesehatan lainnya secara daring [5]. Sementara itu, Alodokter merupakan *platform* layanan kesehatan yang menyediakan layanan konsultasi dokter secara daring, artikel informasi kesehatan, pembelian obat, serta pemesanan layanan kesehatan [6]. Meningkatnya pemanfaatan berbagai aplikasi kesehatan turut mendorong pengguna untuk memberikan penilaian dan umpan balik terhadap suatu layanan melalui *platform* distribusi aplikasi seperti Google Play Store.

Melalui fitur *rating* dan *review* pada Google Play Store memungkinkan pengguna menyampaikan pendapat berdasarkan pengalaman mereka setelah menggunakan aplikasi [7]. Ulasan tersebut menjadi sumber informasi yang bernilai bagi pengembang untuk mengevaluasi kualitas layanan, mengidentifikasi permasalahan yang dihadapi pengguna, serta menentukan prioritas pengembangan aplikasi agar bisa memenuhi kebutuhan pengguna [8]. Namun, jumlah ulasan yang terus bertambah menyebabkan proses analisis secara manual menjadi tidak efisien. Selain itu, ulasan pengguna umumnya ditulis dalam bahasa yang beragam, mengandung singkatan, kata tidak baku, serta kesalahan penulisan, sehingga menyulitkan proses identifikasi sentimen [9]. Oleh karena itu, analisis sentimen menjadi salah satu pendekatan yang banyak dimanfaatkan untuk mengidentifikasi persepsi pengguna secara otomatis berdasarkan ulasan tekstual yang diberikan terhadap suatu aplikasi [8].

Analisis sentimen merupakan teknik komputasi yang menggabungkan *Natural Language Processing*, *data mining*, dan *text mining* untuk mengidentifikasi opini, sikap, emosi, maupun maksud pengguna yang terkandung dalam data berbentuk teks [10], [11]. Keberhasilan analisis sentimen tidak hanya dipengaruhi oleh pemilihan algoritma klasifikasi, tetapi juga oleh kualitas data pelatihan dan proses pelabelan sentimen yang digunakan [12], [13]. Meskipun model berbasis *transformer* diketahui memiliki kemampuan yang lebih baik dalam memahami konteks kalimat yang kompleks dibandingkan algoritma *machine learning* klasik, beberapa penelitian masih berfokus pada pelaporan metrik performa tanpa mengevaluasi karakteristik kesalahan maupun kemampuan generalisasi model pada data di luar distribusi pelatihan [14], [15], [16]. Selain itu, kemampuan model dalam memahami kalimat kontradiktif, yaitu kalimat yang menyajikan dua informasi yang saling bertentangan atau berbeda makna yang umumnya ditandai dengan konjungsi seperti *namun* atau *tetapi* [17], masih jarang dianalisis secara khusus. Oleh karena itu, diperlukan penelitian yang tidak hanya membandingkan performa model, tetapi juga mengkaji model melalui analisis interpretabilitas dan pengujian generalisasi untuk memahami karakteristik serta batas kemampuan model dalam mengklasifikasikan sentimen pada kalimat kontradiktif.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan mengevaluasi kemampuan model *machine learning* klasik dan IndoBERT dalam mengklasifikasikan sentimen ulasan berbahasa Indonesia pada aplikasi layanan kesehatan, khususnya pada kalimat kontradiktif yang mengandung pergeseran sentimen antarklausa. Penelitian ini juga mengevaluasi kualitas metode pelabelan, menguji kemampuan generalisasi model pada skenario *cross-domain* dan *in-domain*, serta menganalisis karakteristik keputusan prediksi IndoBERT menggunakan *Integrated Gradients*. Melalui kerangka evaluasi tersebut, penelitian ini diharapkan memberikan pemahaman yang lebih komprehensif mengenai kemampuan dan keterbatasan kedua model dalam menghadapi kalimat kontradiktif, serta menjadi referensi bagi penelitian selanjutnya dalam pengembangan dan evaluasi model analisis sentimen berbahasa Indonesia.

2 Tinjauan Literatur

Berbagai penelitian telah menerapkan analisis sentimen dengan menggunakan algoritma *machine learning* untuk memahami persepsi pengguna terhadap suatu aplikasi. Fitriyanto *et al.* membandingkan *Support Vector Machine* (SVM), *Logistic Regression*, *K-Nearest Neighbor*, dan *Naïve Bayes* pada ulasan aplikasi dompet digital DANA, dengan hasil *Naïve Bayes* memperoleh akurasi tertinggi sebesar 79,80% dan paling stabil di antara keempat algoritma lainnya [9]. Sementara itu, Kuswati *et al.* menerapkan IndoBERT pada analisis sentimen aplikasi SatuSehat dan mencapai akurasi 86,56% menggunakan 3.000 ulasan [15]. Perbandingan langsung antara model klasik dan *transformer* juga dilakukan oleh Abriananta *et al.* yang menunjukkan bahwa IndoBERT memperoleh akurasi 81,4%, lebih tinggi dibandingkan SVM (74,2%) dan *Naïve Bayes* (72,5%) [16]. Temuan tersebut menunjukkan bahwa model berbasis *transformer* cenderung lebih efektif dalam menangkap konteks bahasa dibandingkan algoritma *machine learning* klasik.

Selain pemilihan algoritma, kualitas pelabelan turut memengaruhi performa model analisis sentimen. Firda *et al.* menunjukkan bahwa pendekatan *InSet Lexicon* menghasilkan akurasi lebih tinggi sebesar 89,7% dibandingkan *rating-based labeling* pada model SVM [18]. Penelitian Rizkia *et al.* juga membandingkan berbagai pendekatan pelabelan, yaitu manual, *InSet Lexicon*, IndoBERT, IndoBERTweet, dan RoBERTa, serta menunjukkan bahwa meskipun IndoBERTweet menghasilkan skor F1 tertinggi, pelabelan manual memberikan distribusi kelas yang lebih representatif [19]. Hasil tersebut menunjukkan bahwa kualitas pelabelan tidak hanya memengaruhi performa model, tetapi juga representativitas data pelatihan.

Dalam kajian linguistik bahasa Indonesia, Tjong *et al.* membandingkan kata penghubung pertentangan dalam bahasa Indonesia dan bahasa Mandarin serta menunjukkan bahwa konjungsi seperti *tetapi* dan *namun* secara konsisten berfungsi merevisi atau membatalkan isi klausa sebelumnya, sehingga klausa setelah konjungsi menjadi penentu makna akhir kalimat [20]. Temuan pada ranah linguistik struktural ini sejalan dengan hasil penelitian Chen *et al.* pada ranah pemrosesan bahasa alami, di mana klasifikasi sentimen berbasis aturan logika mendapati bahwa klausa setelah konjungsi pertentangan seperti *but* umumnya mendominasi orientasi sentimen akhir suatu kalimat dibandingkan klausa sebelumnya [21]. Kedua temuan ini, baik dari ranah linguistik struktural maupun

pemrosesan bahasa alami, secara konsisten menunjukkan pola bahwa dalam konstruksi pertentangan, informasi pada klausa setelah konjungsi memiliki peran yang lebih dominan dalam menentukan interpretasi makna akhir kalimat.

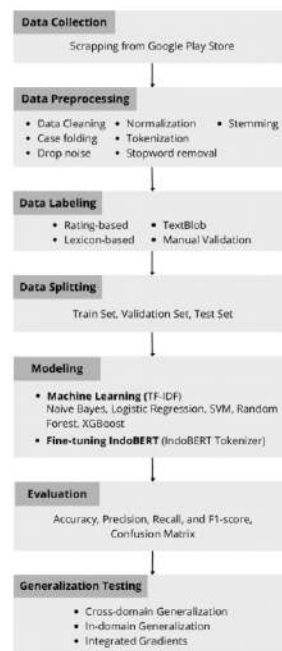
Selain performa dan kualitas pelabelan, kemampuan generalisasi juga menjadi aspek penting dalam mengevaluasi model. Hupkes *et al.* melakukan penelitian mengenai generalisasi pada *natural language processing* dan menjelaskan bahwa kemampuan generalisasi berkaitan dengan kemampuan model untuk menerapkan pengetahuan yang diperoleh dari data sebelumnya pada data dengan karakteristik yang berbeda. Evaluasi yang hanya dilakukan pada data *within-distribution* belum tentu mencerminkan kemampuan model ketika menghadapi perubahan distribusi data. Oleh karena itu, pengujian terhadap data *out-of-distribution* dapat digunakan untuk mengevaluasi kemampuan model dalam mempertahankan performanya ketika terdapat perbedaan distribusi antara data pelatihan dan data pengujian [22].

Interpretabilitas model menjadi aspek penting untuk memahami bagaimana model mengambil keputusan klasifikasi, terutama pada kasus kalimat yang kompleks seperti kalimat kontradiksi. Moraliyage *et al.* menerapkan metode atribusi *Integrated Gradients* untuk mengungkap perilaku model klasifikasi teks dengan mengidentifikasi kata-kata yang memberikan pengaruh positif maupun negatif terhadap prediksi target, sebagai dasar untuk mendeteksi serangan *adversarial* pada *text classifier* [23].

Berdasarkan penelitian terdahulu, dapat disimpulkan bahwa kualitas pelabelan memengaruhi representativitas data pelatihan, sedangkan model berbasis *transformer* umumnya menunjukkan performa yang lebih baik dibandingkan algoritma *machine learning* klasik. Meskipun demikian, sebagian besar penelitian masih berfokus pada pelaporan metrik evaluasi tanpa menjelaskan karakteristik perilaku model ketika menghadapi kalimat kontradiksi yang umum ditemukan pada ulasan pengguna. Selain itu, pengujian generalisasi pada skenario *cross-domain* maupun *in-domain* dengan analisis interpretabilitas masih jarang dilakukan secara bersamaan. Oleh karena itu, penelitian ini mengusulkan kerangka evaluasi yang tidak hanya membandingkan performa *machine learning* klasik dan IndoBERT, akan tetapi juga menganalisis bagaimana model menangani kalimat kontradiksi melalui pengujian generalisasi dan interpretabilitas menggunakan *Integrated Gradients*. Dengan demikian, penelitian ini memberikan pemahaman yang lebih komprehensif mengenai kemampuan serta keterbatasan masing-masing model dalam analisis sentimen ulasan berbahasa Indonesia pada aplikasi layanan kesehatan.

3 Metode Penelitian

Alur penelitian ini terdiri dari beberapa tahapan, dimulai dari pengumpulan data, *preprocessing* data, pelabelan sentimen, pembagian data, pemodelan, evaluasi, hingga pengujian generalisasi model. Gambaran umum alur penelitian disajikan pada Gambar 1.



Gambar 1 Tahapan penelitian

3.1 Data Collection

Penelitian ini menggunakan data ulasan pengguna dari empat aplikasi layanan kesehatan digital di Indonesia yang tersedia di Google Play Store, yaitu Alodokter, Halodoc, Mobile JKN, dan SatuSehat Mobile. Aplikasi ini dipilih karena mewakili domain aplikasi layanan kesehatan digital di Indonesia, baik dari aplikasi layanan publik kesehatan milik pemerintah maupun *platform* layanan kesehatan milik swasta. Pengumpulan data dilakukan menggunakan *library* google-play-scraper dengan mengambil 5.000 ulasan terbaru setiap aplikasi. Ulasan yang dikumpulkan dibatasi pada ulasan yang ditulis dengan Bahasa Indonesia serta pengguna yang berada di wilayah Indonesia.

3.2 Data Preprocessing

Tahapan *data preprocessing* dilakukan untuk mengurangi *noise*, menyeragamkan bentuk kata, serta menghasilkan representasi teks yang sesuai dengan karakteristik model yang digunakan. Tahapan *data preprocessing* pada penelitian ini diawali dengan proses *cleaning* yang meliputi pemilihan kolom yang relevan, penambahan kolom identifikasi aplikasi, penggantian nama kolom, pengecekan *missing values*, penghapusan data duplikat, serta penyaringan ulasan yang terdiri dari kurang dari dua kata. *Dataset* yang telah dibersihkan kemudian digabungkan menjadi satu *dataset* utuh. Setelah proses *cleaning*, dilakukan proses *preprocessing* sebagai berikut:

1. *Case folding* untuk mengubah seluruh karakter pada ulasan menjadi huruf kecil
2. *Drop noise character* untuk menghapus karakter seperti angka, tanda baca, emoji, mention, hashtag, dan URL
3. *Normalization* untuk mengubah kata tidak baku atau *slang* menjadi kata baku
4. *Tokenization* untuk memecah kalimat menjadi token-token kata
5. *Stopword removal* untuk menghilangkan kata sambung atau kata depan yang tidak memberikan informasi tambahan
6. *Stemming* untuk mengubah setiap kata ke bentuk dasarnya

Meskipun demikian, proses *preprocessing* yang diterapkan pada penelitian ini berbeda antara model *machine learning* klasik dan IndoBERT. Seluruh proses *preprocessing* diterapkan pada model klasik karena representasi fiturnya bergantung pada kata-kata hasil ekstraksi fitur. Sebaliknya, model IndoBERT hanya menerapkan dua proses *preprocessing* pertama, yaitu *case folding* dan *drop noise character*. Sedangkan proses *preprocessing* lanjutan, seperti *normalization*, *stopword removal* dan *stemming* tidak diterapkan karena karakteristik model berbasis BERT tidak memerlukan tahapan *preprocessing* yang lebih kompleks, sehingga pemrosesan teks yang mendasar sudah cukup memadai [24]. Adapun untuk proses tokenisasi pada IndoBERT dilakukan menggunakan *tokenizer* bawaan

<http://sistemasi.ftik.unisi.ac.id>

indobert-base-p1 pada tahap penyiapan input model, berbeda dengan *machine learning* klasik yang melakukan tokenisasi saat *preprocessing*.

3.3 Data Labeling

Penelitian ini membandingkan tiga pendekatan pelabelan sentimen, yaitu *rating-based labeling*, *lexicon-based labeling*, dan *TextBlob labeling*. *Rating-based labeling* dilakukan dengan menggunakan pemetaan langsung dari rating yang diberikan pengguna, di mana rating 4-5 dikategorikan sebagai sentimen positif, rating 3 sebagai netral, dan rating 1-2 sebagai negatif. *Lexicon-based labeling* menggunakan kamus InSet *Lexicon* dari Fajri Koto dengan pendekatan akumulasi skor per kata, di mana kalimat dengan skor kumulatif positif dikategorikan positif dan sebaliknya untuk negatif. Sementara itu, *TextBlob labeling* dilakukan dengan menerjemahkan ulasan ke Bahasa Inggris terlebih dahulu menggunakan *library googletrans*, kemudian dihitung polaritas sentimen berdasarkan nilai ambang batas, yaitu polaritas kurang dari 0 sebagai sentimen negatif dan polaritas lebih dari 0 sebagai sentimen positif.

Selanjutnya, ulasan yang telah melewati tahap *data preprocessing* diperiksa dan diberikan label secara manual oleh penulis yang bertindak sebagai anotator tunggal. Label manual tersebut digunakan sebagai acuan (*gold standard*) untuk mengevaluasi hasil pelabelan dari setiap metode. Pemberian label dilakukan berdasarkan makna dan kecenderungan sentimen yang terkandung dalam teks ulasan. Ulasan dikategorikan positif apabila memuat ungkapan kepuasan, pujian, manfaat, pengalaman pengguna yang menyenangkan. Ulasan dikategorikan negatif apabila mengandung keluhan, kritik, ketidakpuasan, dan kendala penggunaan. Sementara label netral diberikan pada ulasan yang hanya menyampaikan informasi, pertanyaan, atau pernyataan di luar konteks penggunaan aplikasi. Khusus pada ulasan kontradiktif, sentimen akhir ditentukan dengan memprioritaskan opini pada klausa yang muncul setelah konjungsi *tapi*, *namun* atau *tetapi* karena klausa tersebut merepresentasikan penekanan opini pengguna.

Hasil pelabelan *rating-based labeling*, *lexicon-based labeling*, dan *TextBlob labeling* kemudian dibandingkan dengan label manual berdasarkan tingkat kesesuaiannya. Metode yang memperoleh tingkat kesesuaian tertinggi akan digunakan sebagai dasar dalam penyusunan label akhir yang selanjutnya disesuaikan kembali berdasarkan hasil validasi manual, terutama pada ulasan yang menunjukkan perbedaan antara hasil pelabelan otomatis dan label manual. Setelah label akhir ditetapkan, ulasan berlabel netral dihapus karena tidak menunjukkan kecenderungan sentimen positif atau negatif dan hanya ulasan berlabel positif dan negatif yang akan digunakan pada tahap pemodelan.

3.4 Data Splitting

Pembagian *dataset* ditentukan berdasarkan kebutuhan proses pelatihan masing-masing model. Pada model *machine learning* klasik, *dataset* dibagi dengan rasio 80:20, yaitu 80% sebagai *training set* dan 20% sebagai *test set*. Sementara itu, IndoBERT menggunakan rasio 80:10:10 yang terdiri atas 80% sebagai *training set*, 10% sebagai *validation set*, dan 10% sebagai *test set*. *Validation set* pada IndoBERT digunakan untuk memantau proses *fine-tuning*, menerapkan *early stopping*, dan memilih *checkpoint* terbaik tanpa melibatkan *test set*, sedangkan *test set* hanya digunakan untuk evaluasi akhir. Berbeda dengan IndoBERT, model klasik tidak menggunakan *validation set* terpisah karena proses pelatihannya dalam penelitian ini tidak melibatkan proses pemantauan, *early stopping*, atau pemilihan *checkpoint* seperti IndoBERT. Dengan demikian, kedua kelompok model menggunakan proporsi *training set* yang sama, sedangkan perbedaannya hanya terletak pada alokasi 20% data lainnya yang digunakan sesuai dengan kebutuhan pelatihan masing-masing model. Untuk proses Pembagian data pada tahap *data splitting* menggunakan fungsi *train_test_split* dari *library* scikit-learn dengan parameter *stratify* berdasarkan label sentimen untuk mempertahankan proporsi setiap kelas.

3.5 Modeling

Pada tahap *modeling*, penelitian ini membandingkan dua jenis model klasifikasi sentimen, yaitu model *machine learning* klasik dan model berbasis *transformer*. Model *machine learning* yang digunakan terdiri dari *Naïve Bayes*, *Logistic Regression*, *Support Vector Machine* (SVM), *Random Forest*, dan *XGBoost*, sedangkan model berbasis *transformer* menggunakan IndoBERT varian indobert-base-p1.

Pada model *machine learning*, teks direpresentasikan ke dalam bentuk numerik menggunakan TF-IDF dengan parameter *ngram_range* = (1,2), *max_features* = 5000, *min_df* = 3, dan *sublinear_tf* = *True*. Representasi fitur tersebut kemudian digunakan sebagai masukan bagi seluruh model *machine learning*. Konfigurasi *hyperparameter* masing-masing model ditunjukkan pada Tabel 1.

Tabel 1 Konfigurasi *hyperparameter* model klasik

Model	Hyperparameter
Naïve Bayes	Default (MultinomialNB)
Logistic Regression	<i>solver</i> =liblinear, <i>max_iter</i> =1000, <i>random_state</i> =42
SVM	<i>kernel</i> =linear, <i>probability</i> =True, <i>random_state</i> =42
Random Forest	<i>n_estimators</i> =200, <i>random_state</i> =42, <i>n_jobs</i> =-1
XGBoost	<i>n_estimators</i> =300, <i>max_depth</i> =6, <i>learning_rate</i> =0.1, <i>subsample</i> =0.8, <i>colsample_bytree</i> =0.8

Berbeda dengan model *machine learning*, IndoBERT tidak memerlukan proses ekstraksi fitur menggunakan TF-IDF. Akan tetapi, teks diproses menggunakan *tokenizer* bawaan IndoBERT untuk menghasilkan representasi *sub-word*, kemudian model di-*fine-tune* menggunakan Hugging Face Trainer dengan konfigurasi *hyperparameter* pada Tabel 2.

Tabel 2 Konfigurasi *hyperparameter* model IndoBERT

Hyperparameter	Nilai
Learning rate	2e-5
Batch size	16
Num train epochs	5
Weight decay	0.01
Warmup ratio	0.1
Random seed	42
Early stopping	2

Selama pelatihan, model dievaluasi pada setiap akhir *epoch*, sedangkan model terbaik dipilih berdasarkan nilai *validation loss* terendah. Selain itu, mekanisme *early stopping* diterapkan untuk menghentikan pelatihan apabila *validation loss* tidak mengalami perbaikan selama dua *epoch* berturut-turut sehingga risiko *overfitting* dapat diminimalkan.

3.6 Evaluation

Performa setiap model dievaluasi menggunakan empat metrik utama yaitu *accuracy*, *precision*, *recall*, dan *F1-score*. *Accuracy* digunakan untuk mengukur ketepatan prediksi secara keseluruhan, *precision* mengukur ketepatan prediksi pada kelas positif, sedangkan *recall* menunjukkan kemampuan model dalam mengidentifikasi seluruh data yang benar-benar positif. Sementara itu, *F1-score* digunakan sebagai ukuran keseimbangan antara *precision* dan *recall*. Nilai *precision*, *recall*, dan *F1-score* dihitung menggunakan *macro average* agar setiap kelas memperoleh bobot yang sama sehingga hasil evaluasi tidak didominasi oleh salah satu kelas sentimen. Selain itu, *confusion matrix* digunakan untuk menganalisis distribusi hasil klasifikasi berdasarkan label aktual dan label prediksi.

3.7 Generalization Testing

Untuk mengevaluasi kemampuan generalisasi model, pengujian dilakukan menggunakan data yang berada di luar distribusi (*out-of-distribution*) dari *dataset* pelatihan. Pengujian dilakukan pada dua skenario, yaitu *cross-domain* dan *in-domain*. Pada skenario *cross-domain*, data uji mengambil ulasan dari empat aplikasi dengan domain berbeda, yaitu Jamsostek Mobile, SIGNAL, Gojek, dan BRImo. Sementara itu, pada skenario *in-domain* mengambil ulasan dari empat aplikasi layanan kesehatan yang tidak termasuk dalam data pelatihan, yaitu KlikDokter, Good Doctor, YesDok, dan MySiloam.

Untuk memperoleh analisis yang lebih komprehensif, data pada setiap skenario dipilih menggunakan metode *purposive sampling* berdasarkan sumber aplikasi dan pola ulasan yang sesuai dengan tujuan penelitian. Setiap skenario menggunakan 80 ulasan yang terdiri atas 24 *short review*, 16 *long review*, dan 40 *contrastive review*. Proporsi *contrastive review* dibuat lebih besar karena kategori tersebut menjadi fokus utama penelitian dan memiliki kompleksitas semantik yang lebih tinggi. Seluruh data uji kemudian diberi label secara manual oleh satu anotator menggunakan ketentuan pelabelan yang sama dengan validasi *dataset* utama. Rancangan ini digunakan sebagai pengujian diagnostik untuk membandingkan kemampuan relatif model pada variasi domain dan pola ulasan yang telah ditetapkan.

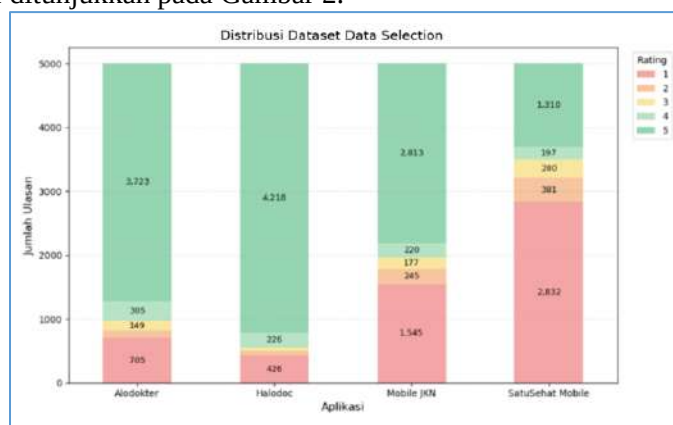
Selain pengujian generalisasi, interpretabilitas IndoBERT pada kategori *contrastive review* dianalisis menggunakan *Integrated Gradients*. Metode ini menghitung atribusi setiap token input terhadap keluaran model dengan mengintegrasikan gradien di sepanjang lintasan dari titik *baseline* menuju input aktual. Besarnya skor atribusi menunjukkan tingkat kontribusi token, sedangkan tanda positif atau negatif menunjukkan arah pengaruhnya terhadap kelas yang diprediksi [25]. Dalam penelitian ini, *Integrated Gradients* diterapkan pada lapisan *embedding* IndoBERT. Tensor *embedding* bernilai nol dengan dimensi yang sama seperti *embedding* input digunakan sebagai *baseline*. Atribusi dihitung melalui 50 *integration steps* dari *baseline* menuju *embedding* input dengan menetapkan kelas hasil prediksi model sebagai kelas target. Skor atribusi setiap token diperoleh dengan menjumlahkan atribusi pada seluruh dimensi *embedding*, kemudian dinormalisasi menggunakan norma L2.

Meskipun data uji telah mencakup beberapa aplikasi dan pola kalimat, interpretasi hasil generalisasi dalam penelitian ini dibatasi pada sampel dan kategori ulasan yang digunakan. Oleh karena itu, hasil tersebut menunjukkan kemampuan relatif model pada rancangan pengujian yang ditetapkan dan tidak secara langsung merepresentasikan seluruh populasi ulasan aplikasi.

4 Hasil dan Pembahasan

4.1 Data Collection

Penelitian ini berhasil mengumpulkan 20.000 ulasan pengguna dari empat aplikasi layanan kesehatan, dengan masing-masing aplikasi berjumlah 5.000 ulasan. Proses pengambilan data dilakukan pada bulan April 2026. Berdasarkan tanggal publikasi ulasan di Google Play Store, ulasan yang terkumpul mencakup periode bulan Maret 2024 hingga April 2026. Distribusi *dataset* hasil proses *data collection* ditunjukkan pada Gambar 2.



Gambar 2 Distribusi *dataset data selection*

Gambar 2 menunjukkan bahwa distribusi pada masing-masing aplikasi memiliki karakteristik yang berbeda. Berdasarkan *rating*, Alodokter dan Halodoc didominasi oleh ulasan dengan *rating* tinggi yang menunjukkan kecenderungan pengguna memberikan penilaian positif terhadap kedua aplikasi tersebut. Sebaliknya, Mobile JKN dan SatuSehat Mobile didominasi oleh ulasan dengan *rating* rendah yang menunjukkan kecenderungan pengguna memberikan penilaian negatif terhadap kedua aplikasi tersebut. Dari distribusi *rating* ini menunjukkan bahwa *dataset* dapat dikatakan seimbang karena sudah mencakup ulasan positif dan ulasan negatif dengan proporsi yang memadai

dan bervariasi. Tabel 3 menyajikan sampel hasil ulasan yang berhasil dikumpulkan dari keempat aplikasi.

Tabel 3 Hasil data selection

Content	Score	At
sangat membantu 🥰	5	2026-04-15 14:52:01
Aplikasi SAMPAH buang buang uang negara.	1	2025-09-11 08:50:53
Sering macet. Tiap buka minta pw. Minta ubah data.		
aplikasi yg bagus dan sangat membantu dalam pembelian obat berikut konsultasi dokter	5	2025-11-26 15:27:26
MANA KODE OTP LEWAT SMS NYAA, udah nungguin dari tadi ga dikirim ² ga becus bgt	1	2026-04-09 17:08:35

4.2 Data Preprocessing

Pada tahap awal dilakukan proses *cleaning* dengan pemeriksaan terhadap kualitas *dataset*. Hasil pengecekan menunjukkan bahwa tidak terdapat *missing values* pada seluruh data ulasan yang digunakan dalam penelitian. Namun, ditemukan sejumlah ulasan duplikat pada keempat aplikasi, yaitu 1.283 ulasan pada Alodokter, 1.496 ulasan pada Halodoc, 1.455 ulasan pada Mobile JKN, dan 616 ulasan pada SatuSehat Mobile. Sebagian besar ulasan duplikat berupa kalimat singkat, seperti “sangat membantu”, “mantap”, “bagus”, “ok”. Untuk menghindari pengaruh tersebut, seluruh ulasan duplikat dihapus dan *dataset* dari keempat aplikasi digabungkan menjadi satu bagian *dataset* dengan melakukan pemilihan kolom yang relevan serta penambahan kolom *app_name* untuk memberikan identitas ulasan berdasarkan asal aplikasinya. Total jumlah data setelah dibersihkan dan digabungkan adalah 15.150 ulasan.

Tahapan selanjutnya adalah proses *preprocessing* teks yang diawali dengan *case folding* untuk mengubah seluruh karakter menjadi huruf kecil, sedangkan *drop noise character* dilakukan untuk menghapus karakter yang tidak memiliki kontribusi informasi terhadap analisis nantinya. Setelah itu, alur *preprocessing* mulai berbeda untuk model *machine learning* klasik dan IndoBERT. Pada model *machine learning* klasik, tahapan *preprocessing* dilanjutkan dengan *normalization*, *tokenization*, *stopword removal*, dan *stemming* sebelum teks dikonversi kembali menjadi *string* untuk proses *feature extraction* TF-IDF. Sebaliknya, model IndoBERT tidak melalui ketiga tahapan tersebut, sehingga ulasan hasil *drop noise character* akan digunakan langsung sebagai *input* pada proses *fine-tuning* IndoBERT. Tabel 4 menyajikan tahapan dan hasil dari proses *preprocessing*.

Tabel 4 Tahapan data preprocessing

Tahapan Preprocessing	Hasil
Ulasan asli	MANA KODE OTP LEWAT SMS NYAA, udah nungguin dari tadi ga dikirim ² ga becus bgt
Case folding	mana kode otp lewat sms nyaa, udah nungguin dari tadi ga dikirim ² ga becus bgt
Drop noise character	mana kode otp lewat sms nyaa udah nungguin dari tadi ga dikirim ga becus bgt
Normalization	mana kode otp lewat sms nya sudah nungguin dari tadi tidak dikirim tidak becus banget
Tokenization	['mana', 'kode', 'otp', 'lewat', 'sms', 'nya', 'sudah', 'nungguin', 'dari', 'tadi', 'tidak', 'dikirim', 'tidak', 'becus', 'banget']
Stopword removal	['mana', 'kode', 'otp', 'lewat', 'sms', 'nya', 'nungguin', 'tadi', 'tidak', 'dikirim', 'tidak', 'becus', 'banget']
Stemming	['mana', 'kode', 'otp', 'lewat', 'sms', 'nya', 'nungguin', 'tadi', 'tidak', 'kirim', 'tidak', 'becus', 'banget']

Total jumlah data setelah dilakukan tahapan *data preprocessing* adalah 15,144 ulasan, untuk distribusi *dataset*-nya ditunjukkan pada Gambar 3.



Gambar 3 Distribusi *dataset* setelah *data preprocessing*

Untuk memberikan gambaran awal mengenai kata-kata yang paling sering muncul pada ulasan aplikasi, dilakukan visualisasi menggunakan *wordcloud* terhadap sentimen positif dan negatif yang ditunjukkan pada Gambar 4.

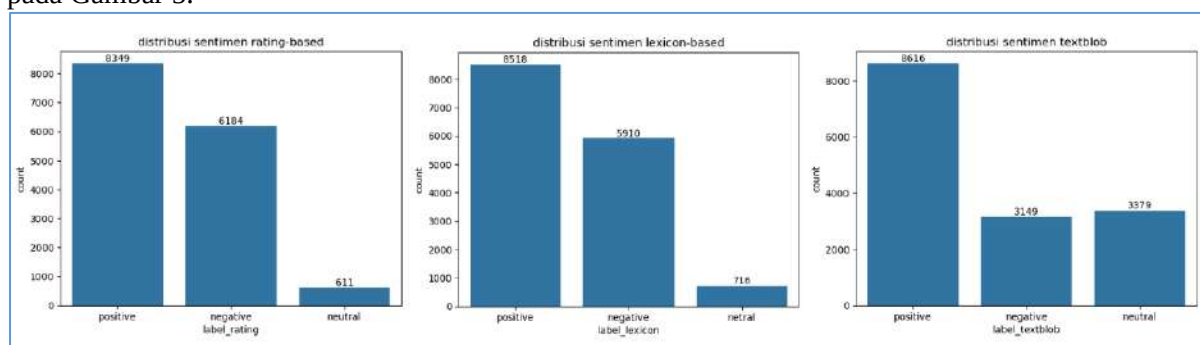


Gambar 4 Visualisasi *wordcloud* positif dan negatif

Hasil visualisasi *wordcloud* menunjukkan bahwa ulasan positif didominasi oleh kata-kata seperti “membantu”, “terima kasih”, “dokter”, “mantap”, dan “pelayanan”. Sementara ulasan negatif didominasi oleh kata-kata “login”, “kode otp”, “aplikasi”, dan “tidak bisa”. Temuan ini menunjukkan bahwa persepsi positif pengguna umumnya berkaitan dengan kualitas layanan aplikasi, sedangkan persepsi negatif lebih banyak dipengaruhi oleh kendala teknis aplikasi.

4.3 Data Labeling

Penelitian ini membandingkan pendekatan *rating-based labeling*, *lexicon-based labeling*, dan *TextBlob labeling*. Distribusi kelas sentimen yang dihasilkan oleh masing-masing metode ditunjukkan pada Gambar 5.



Gambar 5 Distribusi sentimen setiap metode pelabelan

Gambar 5 menunjukkan bahwa ketiga metode menghasilkan distribusi label yang berbeda. *Rating-based labeling* menghasilkan 8.349 ulasan positif, 6.184 ulasan negatif, dan 611 ulasan netral.

<http://sistemasi.ftik.unisi.ac.id>

Lexicon-based labeling memiliki distribusi yang hampir serupa dengan menghasilkan 8.518 ulasan positif, 5.910 ulasan negatif, dan 716 ulasan netral. Sebaliknya untuk *TextBlob labeling* memiliki distribusi yang berbeda, terutama di ulasan netralnya yang meningkat menjadi 3.379 ulasan, 8.616 ulasan positif, dan 3.149 ulasan negatif. Distribusi tersebut menunjukkan bahwa metode *rating-based labeling* dan *lexicon-based labeling* menghasilkan distribusi kelas sentimen yang relatif seimbang, sedangkan untuk metode *TextBlob* cenderung mengklasifikasikan ulasan ke kelas netral.

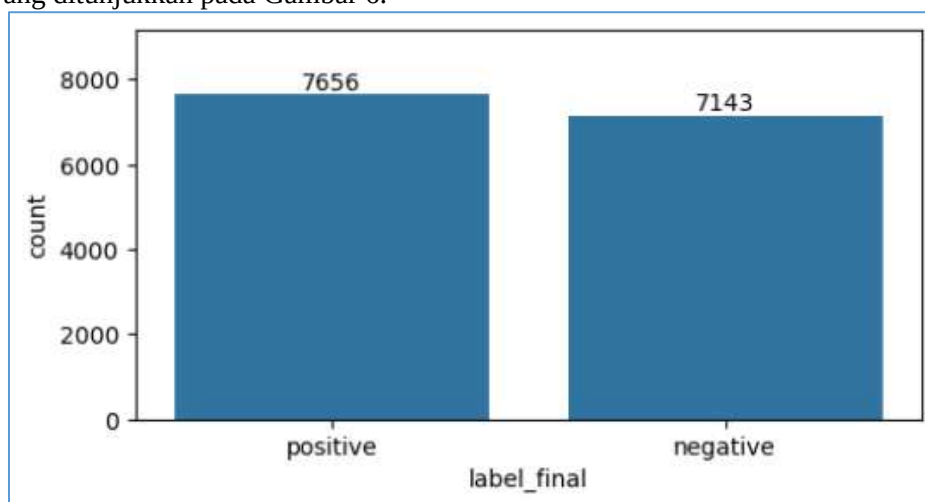
Selanjutnya, ketiga metode tersebut dievaluasi dengan pelabelan manual yang sudah dilakukan dan dijadikan acuan (*gold standard*) untuk mengetahui tingkat kesesuaian pelabelan dengan isi ulasan. Tabel 5 menunjukkan hasil evaluasi dari ketiga metode pelabelan tersebut.

Tabel 5 Hasil evaluasi setiap metode pelabelan

Metode Pelabelan	Tingkat Kesesuaian
<i>Rating-based</i>	90,60%
<i>Lexicon-based</i>	78,65%
<i>TextBlob</i>	63,27%

Hasil evaluasi *rating-based* memiliki tingkat kesesuaian tertinggi daripada dua metode lainnya. Hal ini menunjukkan bahwa *rating* yang diberikan pengguna pada ulasan aplikasi masih selaras dengan isi ulasannya, walau masih ditemui beberapa *rating* ulasan yang diberikan belum selaras dengan isi ulasannya. Kemudian, untuk metode *lexicon-based* masih memiliki keterbatasan dalam memahami konteks kalimat, seperti negasi atau perubahan makna kata akibat kombinasi beberapa kata. Sementara itu, *TextBlob* memiliki akurasi yang paling rendah karena dipengaruhi oleh proses penerjemahan ke Bahasa Inggris, sehingga kalimat mengalami perubahan makna yang berdampak pada hasil klasifikasi kelas sentimen.

Berdasarkan hasil perbandingan tersebut, *rating-based labeling* dipilih sebagai dasar penyusunan label akhir karena memperoleh tingkat kesesuaian tertinggi terhadap pelabelan manual, yaitu 90,60%. Dari 15.144 ulasan yang diperiksa manual, 1.424 label telah diperbaiki dan 13.720 label dipertahankan. Sehingga diperoleh 7.656 ulasan positif, 7.143 ulasan negatif, dan 345 ulasan netral. Setelah ulasan netral dihapus, sebanyak 14.799 ulasan digunakan untuk pemodelan dengan distribusi yang ditunjukkan pada Gambar 6.



Gambar 6 Distribusi dataset data labeling

4.4 Data Splitting

Dataset yang berjumlah 14.799 ulasan dibagi menggunakan dua skema. Pada model *machine learning* klasik, sebanyak 11.839 ulasan digunakan sebagai *train set* dan 2.960 ulasan sebagai *test set*. IndoBERT menggunakan jumlah *train set* yang sama sebanyak 11.839 ulasan, sedangkan 2.960 ulasan sisanya dibagi menjadi *validation set* dan *test set* masing-masing sebanyak 1.480 ulasan. Tabel 6 menunjukkan rincian hasil pembagian untuk kedua model.

Tabel 6 Hasil *train-test split*

Model	Train	Validation	Test	Total
Machine Learning Klasik	11.839 (6.125 Positif; 5.714 Negatif)	-	2.960 (1.531 Positif; 1.429 Negatif)	14.799
IndoBERT	11.839 (6.125 Positif; 5.714 Negatif)	1.480 (766 Positif; 714 Negatif)	1.480 (765 Positif; 715 Negatif)	14.799

4.5 Modeling dan Evaluation

Setelah melalui proses *train-test split*, *train set* dan *test set* yang akan digunakan pada model *machine learning* klasik ditransformasikan menjadi representasi numerik menggunakan TF-IDF. Berdasarkan parameter yang telah ditetapkan, proses ekstraksi menghasilkan 5.000 fitur. Tabel 7 menunjukkan beberapa fitur yang dihasilkan dari proses ekstraksi TF-IDF.

Tabel 7 Fitur hasil ekstraksi TF-IDF

Fitur	Nilai IDF
tidak	1,993066
aplikasi	2,326306
bisa	2,459156
dokter	2,657890
tidak bisa	2,760497

Hasil representasi numerik yang dihasilkan kemudian digunakan sebagai *input* untuk melatih lima algoritma *machine learning* klasik, yaitu *Naïve Bayes*, *Logistic Regression*, *Support Vector Machine* (SVM), *Random Forest*, dan *XGBoost*. Performa masing-masing model kemudian dievaluasi pada *test set* menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Tabel 8 menunjukkan hasil evaluasi dari kelima model klasik.

Tabel 8 Hasil evaluasi model klasik

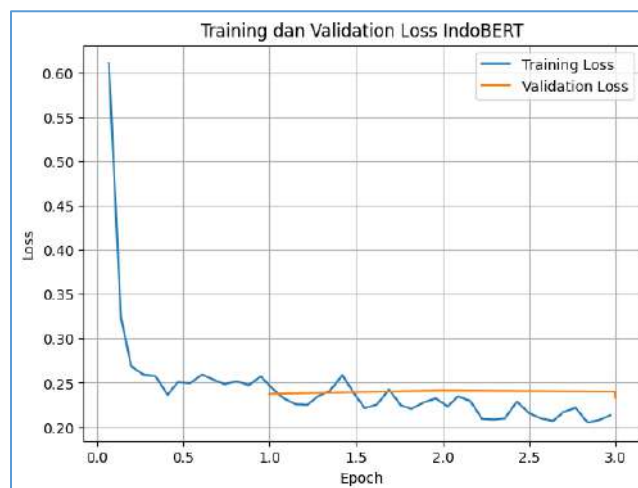
Model	Accuracy	Precision	Recall	F1-score
Naïve Bayes	0,9682	0,9682	0,9682	0,9682
Logistic Regression	0,9753	0,9752	0,9754	0,9753
SVM	0,9757	0,9755	0,9758	0,9757
Random Forest	0,9699	0,9698	0,9701	0,9699
XGBoost	0,9628	0,9629	0,9627	0,9628

Kelima model klasik menghasilkan performa yang sangat baik dengan nilai *F1-score* di atas 96%. SVM menempati posisi dengan performa terbaik, diikuti oleh *Logistic Regression*, *Random Forest*, dan *Naïve Bayes*. Sementara itu, *XGBoost* menghasilkan performa terendah dibandingkan dengan model lainnya. Kemudian nilai *precision* dan *recall* yang seimbang pada setiap model menunjukkan bahwa model tidak cenderung bias ke salah satu kelas. Hal ini diperkuat dengan hasil *confusion matrix* kelima model yang menunjukkan jumlah *true positive* dan *true negative* yang tinggi serta seimbang, sedangkan *false positive* dan *false negative* relatif rendah seperti yang ditunjukkan pada Tabel 9.

Tabel 9 Confusion matrix model *machine learning* klasik

Model	True Positive	True Negative	False Positive	False Negative
Naïve Bayes	1487	1379	50	44
Logistic Regression	1490	1398	31	41
SVM	1491	1401	28	40
Random Forest	1471	1389	40	60

Selanjutnya untuk model IndoBERT di-*fine-tune* menggunakan *dataset* hasil *preprocessing* tanpa proses *normalization*, *stopword removal*, *stemming*, dan *tokenization*. *Tokenization* dilakukan menggunakan *AutoTokenizer* dari *pre-trained* *indobert-base-p1* dengan mengubah kalimat ulasan menjadi token *input* BERT dengan *max_length* 128. Proses pelatihan IndoBERT digambarkan pada kurva *training loss* dan *validation loss* yang ditunjukkan pada Gambar 7.



Gambar 7 Kurva *training loss* dan *validation loss*

Berdasarkan Gambar 7, *validation loss* terendah dicapai pada *epoch* ke-1 dengan nilai 0,2374. Pada *epoch* ke-2, *validation loss* justru naik menjadi 0,2413 dan pada *epoch* ke-3 nilai tersebut turun menjadi 0,2399 dan tetap lebih tinggi dibandingkan hasil *epoch* ke-1. Sementara itu, *training loss* menunjukkan penurunan selama proses pelatihan meskipun terdapat fluktuasi kecil pada beberapa *epoch*. Pola tersebut mengindikasikan mulai munculnya gejala *overfitting* ringan setelah *epoch* ke-1 yang ditandai dengan *training loss* terus menurun sementara *validation loss* tidak lagi mengalami perbaikan. Karena *validation loss* tidak menunjukkan perbaikan selama dua *epoch* berturut-turut (*epoch* ke-2 dan ke-3), mekanisme *early stopping* menghentikan proses pelatihan pada *epoch* ke-3 sebelum mencapai batas maksimal 5 *epoch* yang telah ditentukan. Tabel 10 menunjukkan rincian nilai *eval_loss* pada setiap *epoch*.

Tabel 10 Nilai *eval loss* per *epoch*

<i>Epoch</i>	<i>Validation Loss</i>
1.0	0,2374
2.0	0,2413
3.0	0,2399

Model terbaik yang dihasilkan dari *epoch 1* kemudian dimuat kembali dan dievaluasi ulang, menghasilkan nilai *eval_loss* sebesar 0,2374 yang identik dengan *epoch 1*. Selanjutnya, model tersebut dievaluasi pada *test set* untuk memperoleh hasil evaluasi akhir. Tabel 11 menunjukkan hasil evaluasi dari model IndoBERT.

Tabel 11 Hasil evaluasi model IndoBERT

Model	Accuracy	Precision	Recall	F1-score
IndoBERT	0,9811	0,9814	0,9808	0,9810

Model IndoBERT memperoleh nilai *precision* dan *recall* dengan selisih yang kecil sehingga mengindikasikan model memiliki kemampuan yang seimbang dalam mengenali kedua kelas sentimen dan tidak cenderung bias ke salah satu kelas saja. Hal ini didukung dengan hasil *confusion matrix*

yang menunjukkan nilai *true positive* dan *true negative* yang tinggi serta seimbang dengan *false positive* dan *false negative* yang relatif kecil seperti yang ditunjukkan pada Tabel 12.

Tabel 12 Confusion matrix model IndoBERT

Model	True Positive	True Negative	False Positive	False Negative
IndoBERT	756	696	19	9

Hasil evaluasi perbandingan menunjukkan bahwa IndoBERT memperoleh performa tertinggi dibandingkan seluruh model *machine learning* klasik pada semua metrik evaluasi dengan *accuracy* sebesar 0,9811 dan *F1-score* 0,9810. Dibandingkan SVM sebagai model klasik terbaik, IndoBERT memiliki selisih *accuracy* sebesar 0,0054, *precision* sebesar 0,0059, *recall* sebesar 0,0050, dan *F1-score* sebesar 0,0053 yang lebih tinggi daripada SVM. Keunggulan tersebut diduga disebabkan oleh kemampuan *contextual embedding* pada IndoBERT yang mampu menghasilkan pemahaman semantik yang lebih baik dibandingkan TF-IDF, sehingga model lebih efektif dalam membedakan sentimen pada ulasan aplikasi layanan kesehatan. Tabel 13 menunjukkan perbandingan performa evaluasi seluruh model.

Tabel 13 Perbandingan evaluasi seluruh model

Model	Accuracy	Precision	Recall	F1-score
Naïve Bayes	0,9682	0,9682	0,9682	0,9682
Logistic Regression	0,9753	0,9752	0,9754	0,9753
SVM	0,9757	0,9755	0,9758	0,9757
Random Forest	0,9699	0,9698	0,9701	0,9699
XGBoost	0,9628	0,9629	0,9627	0,9628
IndoBERT	0,9811	0,9814	0,9808	0,9810

Nilai *F1-score* yang tinggi pada seluruh model didukung oleh distribusi *dataset* yang relatif seimbang, yaitu 51,73% ulasan positif dan 48,27% ulasan negatif. Proporsi tersebut dipertahankan pada setiap bagian data dengan menggunakan *stratified splitting*. Potensi *data leakage* dicegah dengan menghapus data duplikat sebelum pembagian data serta memisahkan data pelatihan dan pengujian. Pada model klasik, TF-IDF hanya dipelajari dari *training set*, kemudian digunakan untuk mentransformasi *test set*. Sementara itu, proses *fine-tuning* IndoBERT dilakukan menggunakan *training set*, sedangkan untuk pemilihan *checkpoint* dan *early stopping* IndoBERT menggunakan *validation set*, dan *test set* hanya digunakan pada evaluasi akhir. Selain itu, penggunaan *macro average* pada *precision*, *recall*, dan *F1-score* serta rendahnya *false positive* dan *false negative* pada *confusion matrix* menunjukkan bahwa performa model tidak didominasi oleh salah satu kelas sentimen.

4.6 Generalization Testing

Meskipun selisih performa evaluasi *F1-score* IndoBERT dan SVM sebagai model *machine learning* klasik dengan performa terbaik pada penelitian ini kurang lebih hanya sebesar 0,0053, perbedaan tersebut belum sepenuhnya mencerminkan kemampuan generalisasi model pada data di luar distribusi pelatihan. Oleh karena itu, dilakukan pengujian *out-of-distribution* menggunakan data baru untuk mengevaluasi kemampuan kedua model dalam menghadapi karakteristik ulasan yang lebih beragam [22].

Untuk mengevaluasi kemampuan generalisasi model, dilakukan pengujian *out-of-distribution* menggunakan 80 ulasan baru untuk masing-masing dua skenario, yaitu domain yang berbeda (*cross-domain*) dan domain yang sama (*in-domain*). Tabel 14 menunjukkan hasil perbandingan akurasi seluruh model pada kedua skenario.

Tabel 14 Hasil perbandingan akurasi seluruh model

Model	Cross-Domain	In-Domain	Selisih
Naïve Bayes	85,00%	82,50%	-2,50%
Logistic Regression	87,50%	83,75%	-3,75%

SVM	86,25%	85,00%	-1,25%
Random Forest	81,25%	83,75%	+2,50%
XGBoost	86,25%	82,50%	-3,75%
IndoBERT	88,75%	96,25%	+7,50%

Hasil perbandingan menunjukkan bahwa empat dari lima model *machine learning* klasik, yaitu *Naïve Bayes*, *Logistic Regression*, SVM, dan XGBoost mengalami penurunan akurasi ketika diuji pada skenario *in-domain* dengan penurunan berkisar antara 1,25%-3,75% dibandingkan pada skenario *cross-domain*. Sebaliknya, *Random Forest* justru mengalami kenaikan akurasi sebesar 2,50%, yang menunjukkan bahwa kedekatan domain tidak selalu menghasilkan peningkatan performa pada model *machine learning* klasik. Berbeda dengan model klasik, IndoBERT mengalami peningkatan akurasi dari 88,75% pada skenario *cross-domain* menjadi 96,25% pada skenario *in-domain* atau meningkat sebesar 7,5%. Hasil tersebut menunjukkan bahwa IndoBERT mampu mempertahankan pola generalisasi yang lebih konsisten, di mana performa akurasinya meningkat ketika data uji berasal dari domain yang lebih dekat dengan data pelatihan. Untuk memperoleh pemahaman yang lebih mendalam mengenai karakteristik generalisasi kedua model, analisis akurasi selanjutnya dilakukan berdasarkan kategori kalimat pada masing-masing skenario pengujian. Tabel 15 menunjukkan perbandingan akurasi per kategori kalimat pada kedua skenario.

Tabel 15 Hasil perbandingan akurasi per pola kalimat

Model	Cross-Domain			In-Domain		
	Short Review	Long Review	Contrastive Review	Short Review	Long Review	Contrastive Review
<i>Naïve Bayes</i>	100% (24/24)	100% (16/16)	70,00% (28/40)	95,83% (23/24)	100% (16/16)	67,50% (27/40)
<i>Logistic Regression</i>	100% (24/24)	100% (16/16)	75,00% (30/40)	100% (24/24)	100% (16/16)	67,50% (27/40)
SVM	100% (24/24)	100% (16/16)	72,50% (29/40)	100% (24/24)	100% (16/16)	70,00% (28/40)
<i>Random Forest</i>	95,83% (23/24)	93,75% (15/16)	67,50% (27/40)	95,83% (23/24)	93,75% (15/16)	72,50% (29/40)
XGBoost	95,83% (23/24)	100% (16/16)	75,00% (30/40)	100% (24/24)	100% (16/16)	65,00% (26/40)
IndoBERT	100% (24/24)	100% (16/16)	77,50% (31/40)	100% (24/24)	100% (16/16)	92,50% (37/40)

Baik model *machine learning* klasik maupun IndoBERT memperoleh akurasi yang hampir sempurna pada kategori *short review* dan *long review* di kedua skenario, tetapi menunjukkan perbedaan yang jelas pada *contrastive review*. Pada skenario *cross-domain*, model klasik memperoleh akurasi 67,50%-75,00%, sedangkan IndoBERT lebih tinggi dengan memperoleh 77,50%. Kemudian, pada skenario *in-domain* model klasik memperoleh akurasi 65,00%-72,50%, sementara IndoBERT mencapai 92,50%. Dari total 80 *contrastive review* pada kedua skenario, *Naïve Bayes* menghasilkan 25 kesalahan, *Logistic Regression* dan SVM masing-masing 23 kesalahan, *Random Forest* dan XGBoost masing-masing 24 kesalahan, sedangkan IndoBERT hanya menghasilkan 12 kesalahan. Dibandingkan dengan *Logistic Regression* dan SVM sebagai model klasik dengan kesalahan paling sedikit, IndoBERT mampu mempertahankan performa yang lebih konsisten pada kalimat kontradiktif terutama ketika data uji berasal dari domain yang serupa dengan data pelatihan. Kemampuan tersebut diduga berasal dari *contextual embedding* yang mempertimbangkan hubungan antarkata dan antarklausa, sedangkan TF-IDF pada *machine learning* klasik lebih bergantung pada bobot kemunculan kata sehingga kurang sensitif terhadap perubahan konteks dan pergeseran sentimen dalam satu kalimat.

Analisis lebih lanjut terhadap kesalahan prediksi model *machine learning* klasik menunjukkan bahwa mayoritas kesalahan terjadi pada kategori *contrastive review*, sedangkan *short review* dan *long review* hampir seluruhnya diprediksi dengan benar. Temuan ini mengindikasikan bahwa representasi

TF-IDF yang berbasis *bag-of-words* mengutamakan akumulasi bobot kemunculan kata daripada hubungan antarklausa. Akibatnya, model cenderung mempertahankan prediksi sesuai klausa yang memiliki bobot sentimen lebih dominan, tanpa mampu memahami bahwa klausa setelah konjungsi kontradiktif seperti *tapi*, *tetapi*, dan *namun* sering kali menjadi penentu sentimen akhir kalimat. Kondisi tersebut terlihat pada ulasan “sdh bagus, tapi terkadang cetak dokumen di mesin sering terkendala” yang diprediksi positif karena bobot kata *bagus* lebih dominan dibandingkan kata *terkendala*, meskipun sentimen keseluruhan bersifat negatif. Sebaliknya, ketika klausa kedua memuat kata negatif yang lebih eksplisit, seperti “gak bisa akses” atau “sering error”, model mampu menghasilkan prediksi dengan benar. Temuan ini menunjukkan bahwa keterbatasan model klasik tidak terletak pada perbedaan domain data, melainkan pada ketidakmampuannya mempertahankan konteks ketika terjadi pergeseran sentimen antarklausa. Meskipun demikian, performa model tetap kompetitif pada *short review* dan *long review* menunjukkan bahwa pendekatan TF-IDF yang ditopang oleh tahapan *data preprocessing* tetap efektif untuk kalimat bersentimen tunggal.

Berbeda dengan model *machine learning* klasik, prediksi IndoBERT pada kategori *contrastive review* dianalisis lebih lanjut menggunakan metode *Integrated Gradients* terhadap sampel ulasan yang diprediksi benar maupun salah pada kedua skenario pengujian untuk mengetahui kata atau token mana yang memberikan atribusi atau pengaruh pada keputusan model. Tabel 16 menunjukkan lima sampel ulasan beserta token dan atribusinya.

Tabel 16 Contoh ulasan prediksi IndoBERT

Ulasan	Ground Truth / Prediksi	Token Atribusi	Interpretasi
“...padahal aplikasi jmo sangat bagus & cepat, sangat bermanfaat banget, tpi di download lgi ngabisa dibuka” (<i>cross-domain</i>)	Negatif / Positif (74,45%)	"tpi" (-0,56), "padahal" (-0,51), "bermanfaat" (+0,27), "cepat" (+0,24), "ngab" (+0,04), "##isa" (-0,15)	Konjungsi kontradiktif memperoleh atribusi yang cukup besar, tetapi token pada klausa keluhan <i>ngabisa</i> terfragmentasi menjadi beberapa <i>subword</i> dengan atribusi mengarah ke sentimen positif.
"Sangat membantu saya, respon dokter pun cepat. Namun apl cukup bermasalah, sy sulit menambahkan no hp dan mengirim gambar..." (<i>in-domain</i>)	Negatif / Positif (89,13%)	"membantu" (+0,65), "sangat" (+0,59), "dokter" (+0,25), "namun" (-0,06), "bermasalah" (-0,03), "sulit" (-0,09)	Token pada klausa pembuka memperoleh atribusi lebih tinggi, sedangkan konjungsi <i>namun</i> dan token keluhan pada klausa kedua memberikan kontribusi yang lebih kecil terhadap prediksi.
“Pertama konsultasi tidak dijawab oleh dokternya ;(tapi dapat refund dari klik dokter. Tim CS sangat fast respon terimakasih” (<i>in-domain</i>)	Positif / Negatif (79,17%)	“tidak” (+0,86), “tapi” (-0,12), “ref” (+0,02), “##und” (+0,01), “fast” (-0,28), “terimakasih” (0,32)	Token negasi pada klausa pembuka memperoleh atribusi tertinggi, sedangkan <i>refund</i> terpecah sehingga arah atribusinya berubah.
"aplikasi nya bagus, namun ada kendala yang cukup serius, saya setiap mau pengkinian data untuk pengurusan jmo selalu gagal..." (<i>cross-domain</i>)	Negatif / Negatif (93,89%)	"bagus" (-0,35), "namun" (+0,48), "kendala" (+0,18), "gagal" (+0,21),	Token keluhan setelah <i>namun</i> memberikan kontribusi lebih besar terhadap keputusan model dibandingkan token pujian pada klausa pembuka.
"Apk bagus bangt dan	Negatif /	"apk" (+0,40),	Token <i>tida</i> dan <i>kenapa</i> pada

sangat bermanfaat Tapi kenapa sekarang saya tidak bisa lagi chat n y" (<i>in-domain</i>)	Negatif (93,10%)	"bagus" (-0,05), "bermanfaat" (-0,11), "tapi" (+0,33), "kenapa" (+0,41), "tidak" (+0,52)	klausa kedua memperoleh atribusi tertinggi terhadap prediksi.
--	------------------	--	---

Berdasarkan lima sampel ulasan pada Tabel 16, terlihat bahwa konjungsi kontrasif seperti *tapi* dan *namun* tetap memberikan kontribusi terhadap keputusan model, meskipun besarnya atribusi berbeda pada setiap ulasan. Pada beberapa prediksi yang salah, kontribusi model cenderung lebih didominasi oleh token bermuatan sentimen pada klausa pembuka, sementara token pada klausa berikutnya yang mengubah orientasi sentimen memperoleh atribusi yang lebih rendah. Sebaliknya, pada prediksi yang benar token pada klausa yang merepresentasikan sentimen akhir, seperti *gagal*, *kendala*, dan *tida* memperoleh atribusi yang relatif lebih kuat terhadap keputusan model.

Selain itu, beberapa kata tidak baku atau kata serapan, seperti *ngabisa* dan *refund* terfragmentasi menjadi beberapa *subword* oleh *tokenizer* IndoBERT. Tidak diterapkannya proses *normalization* memungkinkan bentuk kata-kata dengan pola tersebut tetap dipertahankan sebelum tokenisasi sehingga berpotensi menyebabkan informasi terpecah menjadi beberapa *subword* seperti *##isa* serta *ref* dan *##und*. Fragmentasi ini menyebabkan informasi semantik dan nilai atribusi suatu kata tersebar pada beberapa token sehingga kontribusi kata pembawa sentimen menjadi lebih lemah atau kurang konsisten. Kondisi tersebut dapat menurunkan ketepatan prediksi IndoBERT, terutama ketika kata yang terfragmentasi berada pada klausa yang menentukan sentimen akhir. Sebaliknya, *normalization* dapat mengubah kata tidak baku menjadi bentuk yang lebih baku dan mudah dikenali oleh *tokenizer*, mengurangi fragmentasi *subword*, dan menghasilkan representasi informasi sentimen yang lebih utuh. Dengan demikian, *normalization* dapat membantu meningkatkan konsistensi prediksi IndoBERT melalui representasi token yang lebih mudah dikenali. Sedangkan fragmentasi *subword* dapat menurunkan ketepatan prediksi karena informasi sentimen tersebar pada beberapa token. Secara umum, token-token yang memperoleh atribusi tinggi dapat dikelompokkan ke dalam beberapa kategori sebagaimana ditunjukkan pada Tabel 17.

Tabel 17 Kategori token atribusi pada prediksi IndoBERT

Kategori Kata	Kata atau Token	Karakteristik
Kata sifat	"bagus", "baik", "lambat", "nyaman", "mudah"	Memperoleh atribusi besar dan menjadi penentu arah sentimen
Kata kerja	"membantu", "menyukai", "diatasi"	Memperoleh atribusi sedang hingga tinggi
Kata negasi	"tidak", "tidak"	Memperoleh atribusi tertinggi di antara seluruh token ketika merujuk langsung pada objek yang direview
Kata intensitas	"sangat", "banget"	Memperkuat atribusi kata sifat yang mengikutinya
Kata konjungsi	"tapi", "tetapi", "namun"	Atribusinya bervariasi dan menjadi penanda pergeseran sentimen

Hasil analisis pada Tabel 17 menunjukkan bahwa IndoBERT cenderung memberikan atribusi tinggi pada token yang membawa informasi semantik utama, seperti kata sifat, kata negasi, kata intensitas, dan kata kerja. Kata sifat dan kata negasi umumnya memberikan kontribusi terbesar terhadap keputusan model, sedangkan kata intensitas memperkuat makna kata yang menyertainya. Sebaliknya, kata kerja cenderung memperoleh atribusi sedang hingga tinggi. Meskipun konjungsi kontrasif menandai pergeseran sentimen secara linguistik, atribusinya tidak selalu tinggi. Temuan ini menunjukkan bahwa keberhasilan IndoBERT pada kalimat kontrasif lebih ditentukan oleh atribusi pada token pembawa informasi sentimen daripada oleh konjungsi kontrasif itu sendiri.

Temuan tersebut menunjukkan bahwa keunggulan IndoBERT pada pengujian generalisasi tidak serta merta menjadikan model berbasis *transformer* selalu lebih unggul dibandingkan *machine*

learning klasik. Model klasik tetap menunjukkan performa yang kompetitif pada kategori *short review* dan *long review*, sedangkan sebagian kesalahan IndoBERT pada *contrastive review* diduga tidak hanya dipengaruhi oleh kompleksitas hubungan antarklausa, tetapi juga oleh belum diterapkannya *preprocessing* normalisasi kata tidak baku sebelum tokenisasi. Oleh karena itu, penerapan strategi *preprocessing* lanjutan berpotensi meningkatkan kemampuan model berbasis *transformer* dalam menangani beragam variasi kalimat. Secara keseluruhan, hasil penelitian menunjukkan bahwa IndoBERT lebih mampu mempertahankan performa pada kalimat kontradiktif dibandingkan model *machine learning* klasik. Meskipun demikian, keberhasilan model tidak hanya dipengaruhi oleh arsitektur yang digunakan, tetapi juga oleh kompleksitas struktur kalimat serta strategi *preprocessing* yang diterapkan sebelum proses pemodelan.

5 Kesimpulan

Berdasarkan hasil penelitian, IndoBERT menunjukkan performa yang lebih baik dibandingkan model *machine learning* klasik dalam mengklasifikasikan sentimen ulasan aplikasi layanan kesehatan berbahasa Indonesia. IndoBERT memperoleh F1-score sebesar 98,10%, lebih tinggi dibandingkan kelima model *machine learning* klasik dengan F1-score berkisar antara 96,28%-97,57%. Keunggulan IndoBERT tersebut tidak hanya terlihat pada hasil evaluasi, tetapi juga pada pengujian generalisasi khususnya terhadap kalimat kontradiktif yang mengandung pergeseran sentimen antarklausa. Pada kategori kalimat tersebut, IndoBERT menunjukkan performa yang lebih konsisten dibandingkan model *machine learning* klasik pada skenario *cross-domain* maupun *in-domain*. Meskipun demikian, seluruh model *machine learning* klasik tetap menunjukkan performa yang kompetitif pada *short review* dan *long review*, tetapi mengalami penurunan ketika menghadapi struktur kalimat yang lebih kompleks karena representasi TF-IDF kurang mampu menangkap hubungan antarklausa. Analisis menggunakan *Integrated Gradients* menunjukkan bahwa keputusan IndoBERT lebih banyak dipengaruhi oleh token yang membawa informasi sentimen utama, seperti kata sifat, kata negasi, kata intensitas, dan kata kerja daripada oleh konjungsi kontradiktif itu sendiri. Selain membandingkan performa model, analisis ulasan juga menunjukkan pola persepsi pengguna terhadap layanan kesehatan digital di Indonesia. Pengguna yang memberikan sentimen positif umumnya mengapresiasi kemudahan penggunaan aplikasi, manfaat layanan, serta pelayanan tenaga medis yang dinilai responsif khususnya pada aplikasi Halodoc dan Alodokter. Sebaliknya, pengguna yang memberikan sentimen negatif lebih banyak mengeluhkan kendala teknis pada proses autentikasi, seperti kegagalan *login* dan kesulitan memperoleh kode OTP yang banyak ditemukan pada aplikasi Mobile JKN dan SatuSehat Mobile. Temuan ini menunjukkan bahwa pengembang perlu memprioritaskan peningkatan keandalan sistem autentikasi sekaligus mempertahankan kualitas layanan yang telah memperoleh respons positif dari pengguna. Penelitian selanjutnya disarankan mengevaluasi pengaruh penerapan *preprocessing* lanjutan pada model berbasis *transformer*, khususnya normalisasi kata tidak baku dan koreksi kesalahan penulisan sebelum tokenisasi, serta memperkaya variasi kalimat kontradiktif pada data pelatihan untuk meningkatkan kemampuan model dalam menangani pergeseran sentimen pada ulasan pengguna.

Referensi

- [1] P. Julia and H. Ikhsan, "Transformasi Digital dalam Sistem Informasi Kesehatan: Dampak terhadap Kualitas Pelayanan medis," *Public Health and Safety International Journal*, Vol. 5, No. 2, pp. 642–653, Oct. 2025, DOI: 10.55642/phasij.v5i02.1245.
- [2] B. Primin and A. P. Wibowo, "Implementasi Aplikasi berbasis *Mobile* untuk Pelayanan Jasa Kesehatan," *Jurnal Informatika: Jurnal Pengembangan IT*, Vol. 8, No. 2, pp. 119–125, May 2023, DOI: 10.30591/jpit.v8i2.5076.
- [3] T. Sugihartono and R. R. C. Putra, "Penerapan Metode *Support Vector Machine* dalam Classifikasi Ulasan Pengguna Aplikasi *Mobile* JKN," *SKANIKA: Sistem Komputer dan Teknik Informatika*, Vol. 7, No. 2, pp. 144–153, Jul. 2024, DOI: 10.36080/skanika.v7i2.3193.
- [4] S. Setianingsih, H. Hendri, B. O. Lubis, A. Sudradjat, I. Carolina, and W. Widiati, "Evaluasi Aplikasi SATUSEHAT dengan Metode *Use Questionnaire* dan IPA," *JATI (Jurnal Mahasiswa Teknik Informatika)*, Vol. 8, No. 3, pp. 2988–2995, May 2024, DOI: 10.36040/jati.v8i3.9416.

- [5] T. G. W. M. Sidabutar and D. Juardi, "Analisis Sentimen Masyarakat terhadap Penggunaan Halodoc sebagai Layanan *Telemedicine* di Indonesia," *Jurnal Informatika dan Teknik Elektro Terapan*, Vol. 13, No. 1, pp. 571–577, Jan. 2025, DOI: 10.23960/jitet.v13i1.5682.
- [6] A. Putri, A. Farogi, and S. Mukaromah, "Pendekatan Model UTAUT2 dalam Menilai Penerimaan Pengguna terhadap Layanan *Telemedicine* Alodokter," *KONSTELASI: Konvergensi Teknologi dan Sistem Informasi*, Vol. 5, No. 1, pp. 141–154, Jun. 2025, DOI: 10.24002/konstelasi.v5i1.11696.
- [7] A. S. Berliana and M. Mustikasari, "Analisis Sentimen pada Ulasan Aplikasi JAKARTA NOTEBOOK di *Google Play* menggunakan Metode *Recurrent Neural Network (RNN)*," *Jurnal Informatika dan Teknik Elektro Terapan*, Vol. 12, No. 3, pp. 3051–3057, Aug. 2024, DOI: 10.23960/jitet.v12i3.5067.
- [8] A. Komarudin and A. M. Hilda, "Analisis Sentimen Ulasan Aplikasi Identitas Kependudukan Digital pada *Play Store* menggunakan Metode *Naïve Bayes*," *Computer Science (CO-SCIENCE)*, Vol. 4, No. 1, pp. 28–36, Jan. 2024, DOI: 10.31294/coscience.v4i1.2955.
- [9] A. D. Fitriyanto and P. Purwanto, "Analisis Sentimen Ulasan DANA dari *Play Store* dengan Metode *SVM*, *Logistic Regression*, *Naive Bayes* dan *KNN*," *Building of Informatics, Technology and Science (BITS)*, Vol. 7, No. 3, pp. 1887–1899, Dec. 2025, DOI: 10.47065/bits.v7i3.8769.
- [10] F. M. Putra, P. W. Hardjita, and D. A. Tyas, "Sentimen Analisis Pengguna Media Sosial berdasarkan Metode Ekstraksi Fitur dan Klasifikasi," *Jurnal Ilmu Komputer*, Vol. 16, No. 2, pp. 88–96, Sep. 2023, DOI: 10.24843/JIK.2023.v16.i02.p02.
- [11] E. Hokijuliandy, H. Napitupulu, and F. Firdaniza, "Analisis Sentimen menggunakan Metode Klasifikasi *Support Vector Machine (SVM)* dan Seleksi Fitur *Chi-Square*," *SisInfo: Jurnal Sistem Informasi dan Informatika*, Vol. 5, No. 2, pp. 40–49, Aug. 2023, DOI: 10.37278/sisinfo.v5i2.670.
- [12] B. Mahendra, Martanto, D. Pratama, A. Faqih, and R. Kurniawan, "Evaluasi Pengaruh Kualitas Data terhadap Performa Model *Machine Learning* menggunakan Pendekatan *Data-Centric AI*," *Jurnal Sistem Informasi dan Teknologi (SINTEK)*, Vol. 6, No. 1, pp. 107–113, Jan. 2026, DOI: 10.56995/sintek.v6i1.211.
- [13] S. Biswas, K. Young, and J. Griffith, "A Comparison of Automatic Labelling Approaches for Sentiment Analysis," in *Proceedings of the 11th International Conference on Data Science, Technology and Applications*, SCITEPRESS - Science and Technology Publications, Jul. 2022, pp. 312–319, DOI: 10.5220/0011265900003269.
- [14] P. Maxmiliano, Y. F. Riti, I. Y. Nugroho, and C. E. C. Juniarto, "Perbandingan Algoritma *Naive Bayes* dan *Bert* untuk Analisis Sentimen Ulasan Produk Shopee berdasarkan Rating dan Atribut Produk (Warna/Kategori)," *Jurnal Media Informatika*, Vol. 6, No. 5, pp. 2552–2565, Sep. 2025, DOI: 10.55338/jumin.v6i5.6761.
- [15] L. Kuswati and B. A. Habsy, "Analisis Sentimen Multi-Kelas Ulasan Aplikasi Satusel menggunakan *Indobert* berbasis *Transformer*," *Jurnal Sosial dan Teknologi (SOSTECH)*, Vol. 6, No. 4, pp. 1540–1551, Apr. 2026, DOI: 10.59188/jurnalsostech.v6i4.32787.
- [16] H. Abriananta, K. Umam, N. C. H. Wibowo, and M. R. Handayani, "Comparison of Naive Bayes, Support Vector Machine, and Indobert Methods for Classifying Public Sentiment towards the MBG Program on Platform X," *Journal of Applied Informatics and Computing*, Vol. 10, No. 3, pp. 2806–2815, Jun. 2026, DOI: 10.30871/jaic.v10i3.12721.
- [17] Paulia, C. Hualangi, and N. Sugiroto, "Kontrasif Konjungsi Transisi namun dan tetapi dalam Bahasa Indonesia dan Que dan Bu Guo dalam Bahasa Mandarin," *LINGUISTIK: Jurnal Bahasa & Sastra*, Vol. 10, No. 2, pp. 171–185, Jun. 2025, DOI: 10.31604/linguistik.v10i2.171-185.
- [18] H. Firda, P. Putra, N. R. Oktadini, P. E. Sevtiyuni, and A. Meiriza, "Comparison of Rating-based and Inset Lexicon-based Labeling in Sentiment Analysis using SVM (Case Study: GoBiz Application Reviews on Google Play Store)," *SISTEMASI*, Vol. 14, No. 2, pp. 516–529, Mar. 2025, DOI: 10.32520/stmsi.v14i2.4795.
- [19] A. S. Rizkia, W. Wufron, and F. F. Roji, "Analisis Sentimen Coretax: Perbandingan Pelabelan Data Manual, Transformers-based, dan Lexicon-based pada Performa IndoBERT," *MALCOM:*

- Indonesian Journal of Machine Learning and Computer Science*, Vol. 5, No. 3, pp. 1037–1048, Jul. 2025, DOI: 10.57152/malcom.v5i3.2151.
- [20] D. K. Tjong, S. Anwar, A. Hualangi, T. Agustina, and Anita, “Kontrastif Kata Penghubung Pertentangan dalam Bahasa Indonesia dan Bahasa Mandarin,” *LINGUISTIK : Jurnal Bahasa dan Sastra*, Vol. 11, No. 2, pp. 414–431, Jun. 2026, DOI: 10.31604/linguistik.v11i2.414-431.
- [21] Z. Chen, H. Sun, H. He, and P. Chen, “*Learning from Noisy Crowd Labels with Logics*,” in *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, IEEE, Apr. 2023, pp. 41–52, DOI: 10.1109/ICDE55515.2023.00011.
- [22] D. Hupkes *et al.*, “A Taxonomy and Review of Generalization Research in NLP,” *Nat. Mach. Intell.*, Vol. 5, No. 10, pp. 1161–1174, Oct. 2023, DOI: 10.1038/s42256-023-00729-y.
- [23] H. Moraliyage, G. Kulawardana, D. De Silva, Z. Issadeen, M. Manic, and S. Katsura, “*Explainable Artificial Intelligence with Integrated Gradients for the Detection of Adversarial Attacks on Text Classifiers*,” *Applied System Innovation*, Vol. 8, No. 1, Feb. 2025, DOI: 10.3390/asi8010017.
- [24] N. K. Singh, K. Ghosh, J. Mahapatra, U. Garain, and A. Senapati, “*HCDIR: End-to-end Hate Context Detection, and Intensity Reduction model for online comments*,” Dec. 2023. DOI: 10.48550/arXiv.2312.13193.
- [25] M. Sundararajan, A. Taly, and Q. Yan, “*Axiomatic Attribution for Deep Networks*,” in *Proceedings of the 34th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2017, pp. 3319–3328.