



KLASIFIKASI *RATING* ULASAN INSTAGRAM BERBAHASA INDONESIA MENGGUNAKAN MODEL LINEAR DAN *ENSEMBLE* POHON DENGAN INTERPRETASI SHAP

RATING CLASSIFICATION OF INDONESIAN-LANGUAGE INSTAGRAM REVIEWS USING LINEAR AND TREE *ENSEMBLE* MODELS WITH SHAP INTERPRETATION

Indri Anjar Kartika Sari¹, Agung Budi Prasetyo², Hudan Studiawan³, Onno W. Purbo⁴

¹Program Studi Magister Ilmu Komunikasi, UPN "Veteran" Jawa Timur, Surabaya, Indonesia

^{2,3}Institut Teknologi Telkom Surabaya (ITTS), Surabaya, Indonesia

⁴Independent Researcher

¹indri.cs2026@ids.ac.id, ²agung@itts.ac.id, ³hudan@its.ac.id, ⁴onno@indo.net.id

Abstrak

Ulasan aplikasi menyediakan data tekstual yang kaya untuk memahami evaluasi pengguna terhadap produk digital. Namun, penggunaan *rating* bintang sebagai proksi label sentimen dapat menghasilkan ketidakselarasan antara skor dan isi teks, sehingga hasil klasifikasi perlu diinterpretasikan secara hati-hati. Penelitian ini mengevaluasi klasifikasi sentimen berbasis *rating* pada 1.000 ulasan aplikasi Instagram berbahasa Indonesia dengan membandingkan Regresi Logistik, *Random Forest*, *XGBoost*, dan *LightGBM*. Model dilatih menggunakan representasi gabungan TF-IDF n-gram dan fitur leksikal/sintaksis sehingga menghasilkan 5.130 fitur. Interpretabilitas model berbasis pohon dianalisis menggunakan SHAP TreeExplainer untuk mengidentifikasi kontribusi fitur pada tingkat global dan lokal. Pada satu *stratified hold-out split* 80:20, akurasi model berada pada rentang 49,5-50,5%, sedangkan *weighted F1* berada pada rentang 41,5-48,3%. Regresi Logistik memperoleh *weighted F1* tertinggi sebesar 48,3%, tetapi hasil ini tidak diperlakukan sebagai bukti keunggulan umum karena evaluasi belum menggunakan *repeated* atau *k-fold cross-validation*. Lima fitur dengan *mean absolute SHAP* terbesar adalah 'kenapa', 'yang', 'bisa', 'bagus', dan 'untuk'. Temuan tersebut menunjukkan bahwa fitur leksikal yang dominan secara frekuensi dapat memengaruhi keputusan model dan bahwa interpretasi SHAP perlu dibaca bersama karakteristik representasi TF-IDF. Kontribusi penelitian terletak pada *benchmarking* yang transparan dan auditabilitas interpretasi model pada korpus ulasan aplikasi berbahasa Indonesia, sekaligus mendokumentasikan keterbatasan label berbasis *rating*, ketidakseimbangan kelas, ukuran sampel, dan ketidakstabilan estimasi akibat *single split*.

Kata kunci: Analisis sentimen, *Explainable Artificial Intelligence*, SHAP, *XGBoost*, *LightGBM*, ulasan aplikasi, bahasa Indonesia

Article History

Received: September 2026

Reviewed: September 2026

Published: October 2026

Plagiarism Checker No.:

237.u873247.fh8374

Copyright : Author

Publish by : Kohesi



This work is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/)

**Abstract**

App reviews provide rich textual evidence for understanding users' evaluations of digital products. However, using star ratings as a proxy for textual sentiment may introduce rating-text mismatch, requiring cautious interpretation of classification results. This study evaluates rating-based sentiment classification on 1,000 Indonesian-language Instagram app reviews by comparing Logistic Regression, Random Forest, XGBoost, and LightGBM. The models use a combined TF-IDF n-gram and lexical/syntactic representation comprising 5,130 features. SHAP TreeExplainer is applied to the tree-based models to examine global and local feature contributions. Under a single 80:20 stratified hold-out split, model accuracy ranges from 49.5% to 50.5%, while weighted F1 ranges from 41.5% to 48.3%. Logistic Regression achieves the highest weighted F1 (48.3%), but this result is not interpreted as evidence of general superiority because repeated or k-fold cross-validation has not been performed. The five features with the largest mean absolute SHAP values are 'kenapa', 'yang', 'bisa', 'bagus', and 'untuk'. These findings indicate that highly frequent lexical features can influence model decisions and that SHAP interpretations should be considered together with the characteristics of TF-IDF representations. The contribution of this study is a transparent and auditable benchmark of classical and tree-based classifiers for Indonesian app reviews, together with explicit reporting of rating-label limitations, class imbalance, sample size, and single-split uncertainty.

Keywords: Sentiment analysis, Explainable Artificial Intelligence, SHAP, XGBoost, LightGBM, app reviews, Indonesian language.

1. PENDAHULUAN**1.1 Latar Belakang**

Indonesia merupakan salah satu pasar aplikasi seluler terbesar di dunia, dengan basis pengguna media sosial yang terus tumbuh dan interaksi digital yang intensif melalui platform berbagi ulasan. Ulasan pengguna (*user review*) pada toko aplikasi bukan sekadar umpan balik konsumen, melainkan juga korpus data bahasa alami yang merefleksikan sikap, keluhan, harapan, dan kepuasan pengguna terhadap sebuah produk digital secara langsung dan tidak terstruktur. Bagi pengembang aplikasi, pemahaman terhadap pola sentimen dalam ulasan ini menjadi krusial untuk perbaikan produk berkelanjutan; bagi peneliti ilmu komunikasi dan ilmu komputer, korpus ini menjadi laboratorium alami untuk menguji metode pemrosesan bahasa alami (*natural language processing*/NLP) pada teks berbahasa Indonesia yang khas dengan campuran bahasa gaul, singkatan, dan struktur kalimat informal.

Perkembangan pembelajaran mesin (*machine learning*) dalam satu dekade terakhir telah mendorong akurasi klasifikasi sentimen ke tingkat yang semakin tinggi, khususnya melalui model *ensemble* berbasis pohon keputusan seperti *Random Forest*, *XGBoost*, dan *LightGBM*, serta model berbasis *transformer* seperti BERT dan turunannya untuk bahasa Indonesia (IndoBERT). Namun demikian, peningkatan akurasi ini kerap kali berbanding terbalik dengan



tingkat keterjelasan (*interpretability*) model: semakin kompleks arsitektur yang digunakan, semakin sulit bagi manusia untuk memahami mengapa suatu prediksi dihasilkan. Persoalan ini dikenal luas dalam literatur sebagai dilema akurasi-interpretabilitas (*accuracy-interpretability trade-off*), dan menjadi pendorong utama lahirnya bidang kecerdasan buatan yang dapat dijelaskan atau *Explainable Artificial Intelligence* (XAI).

Salah satu metode XAI yang paling banyak diadopsi dalam beberapa tahun terakhir adalah SHAP (*SHapley Additive exPlanations*), yang berakar dari teori permainan kooperatif *Shapley value* untuk mengalokasikan kontribusi tiap fitur terhadap *output* model secara adil dan konsisten secara matematis. SHAP memungkinkan peneliti tidak hanya melaporkan metrik performa model, tetapi juga menjelaskan *token* atau fitur spesifik apa yang mendorong sebuah ulasan diklasifikasikan sebagai positif, negatif, atau netral – baik pada tingkat keseluruhan dataset (*global explanation*) maupun pada tingkat satu sampel tunggal (*local explanation*).

Studi ini berangkat dari kebutuhan praktis dan akademis untuk menghadirkan sistem klasifikasi sentimen ulasan aplikasi berbahasa Indonesia yang transparan dan dapat diaudit. Berbeda dari banyak studi yang berfokus semata pada pengejaran angka akurasi tertinggi, penelitian ini secara eksplisit menempatkan audit integritas data dan interpretabilitas model sebagai bagian inti dari kontribusi ilmiah, dengan melaporkan secara jujur keterbatasan data yang dihadapi selama proses penelitian.

1.2 Rumusan Masalah

Penelitian ini berfokus pada perbandingan model linear dan *ensemble* berbasis pohon untuk mengklasifikasikan ulasan aplikasi Instagram berbahasa Indonesia menggunakan label berbasis *rating*. Evaluasi diarahkan pada perbedaan kinerja model dalam representasi fitur yang sama serta kontribusi fitur terhadap keluaran model berbasis pohon. Karena *rating* tidak identik dengan sentimen tekstual, interpretasi hasil dibatasi pada target berbasis *rating* dan tidak digunakan untuk menyimpulkan polaritas teks yang telah tervalidasi melalui anotasi manusia.

1.3 Pertanyaan Penelitian

Kode	Pertanyaan Penelitian
RQ1	Faktor tekstual apa yang paling berkaitan dengan klasifikasi sentimen ulasan aplikasi Instagram berbahasa Indonesia?
RQ2	Bagaimana perbandingan performa antara model linear (Regresi Logistik) dan model <i>ensemble</i> berbasis pohon (<i>Random Forest</i> , <i>XGBoost</i> , <i>LightGBM</i>) pada tugas klasifikasi ini?
RQ3	Fitur atau <i>token</i> apa yang secara konsisten muncul sebagai penentu keputusan utama menurut analisis SHAP, baik secara global maupun lokal?
RQ4	Sejauh mana keterbatasan data (ketiadaan metrik keterlibatan dan dimensi temporal) memengaruhi validitas dan generalisasi temuan penelitian ini?

1.4 Tujuan dan Kontribusi Penelitian

Penelitian ini bertujuan untuk: (1) membangun dan membandingkan empat model klasifikasi sentimen pada korpus ulasan aplikasi Instagram berbahasa Indonesia; (2) menerapkan analisis SHAP untuk mengungkap fitur-fitur penentu keputusan model secara transparan; dan (3) mendokumentasikan secara eksplisit keterbatasan data sebagai bagian dari kontribusi metodologis penelitian. Kontribusi utama penelitian ini terletak pada tiga aspek. Pertama, kontribusi metodologis berupa kerangka audit status data (AVAILABLE/PARTIAL/MISSING) yang dapat direplikasi pada studi *big-data* sosial media lain yang menghadapi keterbatasan akses



data serupa. Kedua, kontribusi empiris berupa perbandingan performa dan interpretabilitas empat algoritme pembelajaran mesin pada korpus ulasan aplikasi berbahasa Indonesia yang relatif jarang diteliti dibandingkan bahasa Inggris. Ketiga, kontribusi praktis berupa identifikasi kosakata penentu sentimen yang dapat dimanfaatkan pengembang aplikasi maupun peneliti linguistik komputasional Indonesia untuk pengembangan leksikon sentimen yang lebih baik.

1.5 Sistematika Penulisan

Artikel ini disusun sebagai berikut. Bagian 2 menyajikan tinjauan pustaka yang mencakup perkembangan analisis sentimen, kecerdasan buatan yang dapat dijelaskan, dan posisi penelitian ini di antara studi-studi terdahulu. Bagian 3 menjelaskan metodologi penelitian secara rinci, mencakup sumber data, praproses, rekayasa fitur, arsitektur model, dan skema evaluasi. Bagian 4 memaparkan hasil eksperimen. Bagian 5 mendiskusikan temuan dalam konteks literatur yang lebih luas. Bagian 6 menguraikan keterbatasan penelitian dan arah pengembangan mendatang. Bagian 7 menutup artikel dengan kesimpulan.

1.6 Ruang Lingkup dan Batasan Awal Penelitian

Ruang lingkup penelitian ini secara khusus dibatasi pada klasifikasi sentimen tingkat dokumen (*document-level sentiment classification*) terhadap ulasan aplikasi Instagram berbahasa Indonesia, dan tidak mencakup analisis sentimen tingkat aspek (*aspect-based sentiment analysis*) yang memerlukan anotasi tambahan pada level entitas atau fitur produk spesifik. Penelitian ini juga tidak mencakup deteksi emosi multi-label (*multi-label emotion detection*) maupun analisis ironi/sarkasme, meskipun kedua bidang tersebut memiliki relevansi tinggi terhadap konteks bahasa Indonesia informal dan diusulkan sebagai agenda penelitian lanjutan pada Bagian 6.2. Batasan lingkup ini ditetapkan secara sengaja pada tahap awal penelitian untuk menjaga kedalaman analisis pada cakupan yang lebih sempit namun tuntas, alih-alih memperluas cakupan penelitian secara dangkal ke berbagai subtugas NLP yang masing-masing memerlukan validasi metodologis tersendiri.

2. TINJAUAN PUSTAKA

2.1 Analisis Sentimen: Perkembangan dan Pendekatan

Analisis sentimen (*sentiment analysis*) merupakan cabang pemrosesan bahasa alami yang bertujuan mengidentifikasi dan mengekstraksi opini, emosi, serta polaritas afektif dari data tekstual. Pendekatan awal dalam bidang ini didominasi oleh metode berbasis leksikon (*lexicon-based*) yang mengandalkan kamus kata bermuatan sentimen serta metode pembelajaran mesin klasik seperti Naïve Bayes dan *Support Vector Machine* yang dilatih di atas representasi *bag-of-words* maupun TF-IDF. Seiring waktu, penelitian bergeser menuju model pembelajaran mendalam (*deep learning*) dan arsitektur *transformer* yang mampu menangkap konteks semantik yang lebih kaya, meskipun konsekuensinya adalah meningkatnya kompleksitas komputasi dan menurunnya transparansi model.

Untuk konteks bahasa Indonesia, tantangan analisis sentimen menjadi lebih kompleks karena karakteristik linguistik yang khas: percampuran bahasa formal dan informal (bahasa gaul), penggunaan singkatan tidak baku, variasi dialek regional, serta pengaruh bahasa daerah dan bahasa asing dalam percakapan sehari-hari di media sosial. Leksikon sentimen berbahasa Inggris tidak dapat langsung diterapkan tanpa adaptasi, sehingga penelitian pada domain ini umumnya memerlukan leksikon sentimen Indonesia yang dikembangkan secara khusus atau model bahasa terlatih seperti IndoBERT yang dilatih atas korpus besar teks Indonesia.

Sumber daya pemrosesan bahasa Indonesia telah tersedia, tetapi kesesuaian domain, variasi bahasa informal, dan definisi label tetap perlu diperhatikan ketika sumber daya tersebut digunakan untuk menganalisis ulasan aplikasi. Oleh karena itu, perbandingan hasil antar penelitian perlu mempertimbangkan karakteristik korpus dan protokol evaluasinya. Pada korpus penelitian ini, *rating* 1 merupakan kelas terbesar, sedangkan *rating* 4 merupakan kelas terkecil. Distribusi tersebut menjadi dasar untuk mengevaluasi kinerja per kelas. Pengaruh

ketidakseimbangan terhadap perbedaan kinerja algoritme belum dapat dipisahkan melalui eksperimen yang dilaporkan.

2.2 Model Ensemble Berbasis Pohon dalam Klasifikasi Teks

Random Forest, sebagai salah satu metode *ensemble* bagging yang diperkenalkan oleh Breiman, membangun sejumlah besar pohon keputusan dari sampel *bootstrap* data dan menggabungkan prediksinya melalui voting mayoritas. Pendekatan ini terbukti tangguh terhadap *overfitting* dibandingkan pohon keputusan tunggal, namun kurang optimal ketika berhadapan dengan ruang fitur berdimensi sangat tinggi seperti representasi TF-IDF pada korpus teks.

XGBoost (*Extreme Gradient Boosting*) dan LightGBM (*Light Gradient Boosting Machine*) merupakan dua varian metode *boosting* yang membangun pohon secara sekuensial, di mana setiap pohon baru berupaya memperbaiki kesalahan residual dari pohon sebelumnya. XGBoost dikenal karena regularisasi eksplisit yang mengurangi risiko *overfitting*, sementara LightGBM mengadopsi strategi pertumbuhan pohon berbasis daun (*leaf-wise growth*) yang secara komputasional lebih efisien pada dataset besar. Kedua algoritme ini telah banyak diadopsi dalam kompetisi ilmu data dan penelitian klasifikasi teks karena kemampuannya menangani fitur berdimensi tinggi dan interaksi non-linear antar fitur, meskipun keduanya tetap tergolong model kotak-hitam yang memerlukan lapisan interpretabilitas tambahan untuk dapat dipahami secara manusiawi.

Selain ketiga algoritme di atas, literatur juga mencatat pendekatan *boosting* lain seperti CatBoost yang dioptimalkan khusus untuk menangani fitur kategorikal, serta pendekatan *stacking ensemble* yang menggabungkan prediksi dari beberapa model dasar melalui model *meta-learner*. Penelitian ini memilih untuk membatasi perbandingan pada empat model (Regresi Logistik, *Random Forest*, XGBoost, LightGBM) dengan pertimbangan bahwa keempatnya merepresentasikan spektrum kompleksitas yang cukup luas dari model linear paling sederhana hingga dua varian *boosting* yang paling banyak diadopsi dalam praktik industri sembari tetap menjaga kelayakan komputasi dan kejelasan interpretasi hasil perbandingan bagi pembaca.

2.3 Explainable Artificial Intelligence dan SHAP

Meningkatnya adopsi model pembelajaran mesin yang kompleks dalam pengambilan keputusan telah memicu kebutuhan mendesak akan metode yang dapat menjelaskan bagaimana dan mengapa sebuah model menghasilkan prediksi tertentu. *Explainable Artificial Intelligence* (XAI) hadir sebagai respons terhadap kebutuhan ini, dengan tujuan meningkatkan kepercayaan, akuntabilitas, dan kemampuan audit terhadap sistem kecerdasan buatan, terutama pada domain yang berdampak langsung terhadap manusia.

SHAP, yang diperkenalkan oleh Lundberg dan Lee, merupakan salah satu kerangka XAI yang paling banyak diadopsi karena landasan teoretisnya yang kuat dalam teori permainan kooperatif (*cooperative game theory*), khususnya konsep *Shapley value* yang menjamin sifat konsistensi dan keadilan alokasi kontribusi antar fitur. Untuk model berbasis pohon seperti *Random Forest*, XGBoost, dan LightGBM, SHAP menyediakan varian TreeExplainer yang secara efisien menghitung nilai Shapley tanpa memerlukan aproksimasi berat komputasi seperti pada KernelSHAP. Studi-studi terdahulu telah menunjukkan penerapan SHAP secara luas pada berbagai domain analisis sentimen, mulai dari ulasan produk daring, opini publik di media sosial, hingga aplikasi klinis dan keuangan, di mana kombinasi antara model *ensemble* berbasis pohon dan SHAP TreeExplainer menjadi pola metodologis yang umum ditemukan.

2.4 Ulasan Aplikasi sebagai Objek Penelitian

Ulasan aplikasi (*app review*) pada dasarnya berbeda dari data media sosial umum karena karakteristiknya yang lebih terarah pada evaluasi fungsional suatu produk digital tertentu, disertai skor bintang (*rating*) yang secara implisit merepresentasikan polaritas sentimen. Penelitian pada domain ini relevan bagi disiplin manajemen produk digital, UX research, dan



linguistik komputasional, karena memungkinkan identifikasi keluhan atau apresiasi pengguna secara granular pada level fitur produk. Namun demikian, ulasan aplikasi juga rentan terhadap ketidakseimbangan distribusi *rating* fenomena yang dikenal sebagai *J-shaped distribution*, di mana ulasan sangat positif (bintang 5) dan sangat negatif (bintang 1) mendominasi dibandingkan *rating* netral sebuah karakteristik yang turut teramati pada dataset penelitian ini.

2.5 Posisi Penelitian dan Kesenjangan Literatur (Research Gap)

Literatur yang ditinjau menunjukkan tiga area yang masih dapat diperkuat dalam studi ini. Pertama, evaluasi pada dataset kecil dan tidak seimbang memerlukan pelaporan metrik yang lebih lengkap daripada *accuracy* dan *weighted F1*. Kedua, penggunaan SHAP pada model berbasis pohon dapat memberikan audit fitur yang tidak tersedia dari metrik agregat saja. Ketiga, *rating* sebagai *weak/proxy* label perlu dibedakan secara eksplisit dari *sentiment annotation* berbasis teks. Penelitian ini memosisikan kontribusinya secara sempit pada kombinasi *benchmarking* empat *classifier*, audit interpretabilitas SHAP, dan dokumentasi keterbatasan data pada korpus ulasan Instagram berbahasa Indonesia.

2.6 Kerangka Konseptual Penelitian

Kerangka konseptual penelitian ini menghubungkan tiga blok utama secara linier: (1) blok representasi data, yang mentransformasikan teks ulasan mentah menjadi representasi numerik melalui kombinasi fitur leksikal-sintaksis dan TF-IDF n-gram; (2) blok pemodelan prediktif, yang membandingkan performa empat algoritme klasifikasi dengan kompleksitas berjenjang, mulai dari model linear hingga *ensemble* berbasis pohon; dan (3) blok interpretabilitas, yang mengungkap kontribusi fitur terhadap keputusan model melalui SHAP TreeExplainer pada level global maupun lokal. Ketiga blok ini tidak diposisikan sebagai tahapan yang terpisah, melainkan sebagai siklus umpan balik konseptual: temuan pada blok interpretabilitas idealnya menginformasikan penyempurnaan pada blok representasi data pada iterasi penelitian berikutnya, sebuah pendekatan yang selaras dengan praktik human-in-the-loop *machine learning* yang semakin banyak diadvokasikan dalam literatur XAI kontemporer.

Kerangka ini secara khusus menolak asumsi bahwa peningkatan akurasi model harus selalu menjadi tujuan tunggal penelitian analisis sentimen. Sebaliknya, penelitian ini memosisikan *trade-off* antara akurasi dan interpretabilitas sebagai objek kajian itu sendiri, bukan sekadar batasan teknis yang perlu diminimalkan. Pendekatan konseptual semacam ini relevan khususnya bagi konteks bahasa Indonesia, di mana ketersediaan korpus berlabel berkualitas tinggi masih terbatas dibandingkan bahasa Inggris, sehingga kemampuan menjelaskan perilaku model pada dataset berukuran kecil menjadi bernilai strategis melebihi sekadar mengejar angka akurasi tertinggi yang mungkin tidak dapat direplikasi pada domain lain.

2.7 Proposisi Penelitian

Berdasarkan tinjauan pustaka di atas, penelitian ini merumuskan tiga proposisi kerja yang diuji secara eksploratif (bukan hipotesis statistik formal, mengingat sifat penelitian yang deskriptif-komparatif): Proposisi 1 (P1), model *ensemble* berbasis pohon akan menunjukkan performa akurasi yang setara atau lebih rendah dibandingkan model linear pada dataset berukuran kecil dengan dimensi fitur tinggi, akibat risiko *overfitting* yang lebih besar pada rasio sampel-terhadap-fitur yang rendah. Proposisi 2 (P2), fitur-fitur yang teridentifikasi sebagai signifikan secara global melalui SHAP akan mencerminkan kombinasi antara kosakata bermuatan sentimen eksplisit dan pola struktural kalimat (seperti kata tanya retorik) yang khas pada bahasa Indonesia informal. Proposisi 3 (P3), ketidakseimbangan distribusi kelas akan menjadi determinan performa model yang lebih dominan dibandingkan pilihan algoritme itu sendiri. Ketiga proposisi ini diuji kembali terhadap temuan empiris pada Bagian 4 dan didiskusikan secara kritis pada Bagian 5.



2.8 Sintesis Tinjauan Pustaka

Tabel berikut merangkum posisi penelitian ini relatif terhadap dimensi utama yang telah dibahas pada Bagian 2.1 hingga 2.5, sebagai jembatan menuju penjelasan metodologi pada Bagian 3.

Tabel 0. Sintesis Dimensi Tinjauan Pustaka dan Posisi Penelitian

Dimensi	Kondisi Umum pada Literatur	Posisi Penelitian Ini
Metrik pelaporan	Berfokus pada akurasi/F1 semata	Akurasi, F1, dan interpretabilitas SHAP
Bahasa	Dominan bahasa Inggris	Bahasa Indonesia informal (ulasan aplikasi)
Model	Sering satu model tunggal	Perbandingan 4 model berjenjang kompleksitas
Pelaporan keterbatasan data	Jarang eksplisit/terstruktur	Status AVAILABLE/PARTIAL/MISSING eksplisit
Reproduktibilitas	Bervariasi, kode tidak selalu terbuka	Kode, data, dan model terlatih tersedia terbuka

3. METODOLOGI PENELITIAN

3.1 Desain Penelitian

Penelitian ini menggunakan desain kuantitatif dengan pendekatan eksperimental komparatif, di mana empat algoritme pembelajaran mesin dilatih dan dievaluasi pada dataset yang identik untuk memungkinkan perbandingan performa yang adil (*controlled comparative benchmarking*). Keseluruhan *pipeline* penelitian dirancang sebagai rangkaian tahap yang dapat diaudit dan direproduksi, mulailah dari pembersihan data mentah hingga interpretasi model akhir melalui SHAP. Setiap tahap *pipeline* menghasilkan artefak audit (file log dan laporan JSON) yang mencatat status ketersediaan data pada tahap tersebut, sesuai prinsip transparansi yang dijunjung penelitian ini.

3.2 Sumber dan Karakteristik Data

Definisi label. Dalam penelitian ini, *rating* bintang digunakan sebagai *weak/proxy* label untuk kelas sentimen. Dengan demikian, istilah 'sentimen' dalam hasil klasifikasi merujuk pada kelas yang diturunkan dari *rating*, bukan *gold-standard annotation* atas polaritas teks. Perbedaan ini penting karena sebuah ulasan dapat mengandung polaritas tekstual yang tidak sepenuhnya konsisten dengan *rating*. Studi khusus mengenai ketidakkonsistenan *review-rating* pada platform digital juga menunjukkan bahwa kedua sinyal tersebut tidak selalu identik.

Penelitian menggunakan 1.000 ulasan aplikasi Instagram berbahasa Indonesia yang tersimpan dalam berkas *Review Instagram.csv*. Kolom yang tersedia adalah *Username*, *Review Text*, dan *Rating*. Berdasarkan audit yang dilaporkan, tidak terdapat nilai kosong atau duplikasi teks, dan terdapat 992 nama pengguna unik. Data dibagi menggunakan stratifikasi berdasarkan kelas target dengan rasio latih-uji 80:20. Keempat *classifier* dibandingkan menggunakan representasi fitur yang sama. Tabel 1 merangkum karakteristik utama dataset.



Tabel 1. Ringkasan Karakteristik Dataset

Metrik	Nilai	Status Verifikasi
Berkas sumber	Review Instagram.csv	Terverifikasi
Jumlah baris (ulasan)	1.000	Terverifikasi
Skema kolom	<i>Username</i> , Review Text, <i>Rating</i>	Terverifikasi
Nilai kosong (missing values)	0	Terverifikasi
Baris duplikat pada teks	0	Terverifikasi
Penulis unik	992	Terverifikasi

Tabel 2. Distribusi Kelas *Rating*

<i>Rating</i>	Jumlah Ulasan	Proporsi
1	481	48,1%
2	156	15,6%
3	134	13,4%
4	82	8,2%
5	147	14,7%

Sebagaimana tampak pada Tabel 2, distribusi kelas *rating* sangat timpang, dengan *rating* 1 mendominasi hampir separuh dataset (48,1%), sementara *rating* 4 menjadi kelas minoritas (8,2%). Ketidakseimbangan ini menjadi faktor signifikan yang memengaruhi performa model klasifikasi dan didiskusikan lebih lanjut pada Bagian 5.

3.3 Pernyataan Keterbatasan Data (*Data Limitation Notice*)

Sejalan dengan prinsip integritas ilmiah yang dijunjung penelitian ini, tiga keterbatasan data dinyatakan secara eksplisit sejak tahap awal penelitian, alih-alih ditemukan setelah eksperimen berjalan. Pertama, dari sisi modalitas data, dataset yang tersedia merupakan ulasan konsumen terhadap aplikasi (*app store reviews*) dan bukan data unggahan atau keterangan (*caption*) konten publik Instagram. Kedua, metrik keterlibatan pada level unggahan (*likes*, *comments*, *shares*, *saves*, *impressions*, *views*) tidak tersedia dalam dataset ini, sehingga metrik skor viralitas asli (*native Viral Score*) yang semula direncanakan dalam desain penelitian awal berstatus PARTIAL dan tidak dapat dihitung secara valid pada tahap ini. Ketiga, dimensi temporal berupa cap waktu pembuatan ulasan yang kontinu tidak tersedia, sehingga analisis deret waktu (*time-series*) berstatus PARTIAL/MISSING, dan penelitian ini secara tegas tidak menggunakan tanggal simulasi (*simulated dates*) sebagai pengganti data yang hilang, karena praktik tersebut berpotensi menghasilkan kesimpulan yang menyesatkan.

Sebagai konsekuensi dari keterbatasan ini, penelitian dirumuskan ulang dari fokus awal prediksi viralitas konten menjadi klasifikasi sentimen ulasan aplikasi yang dapat dijelaskan sebuah reorientasi yang memastikan setiap klaim ilmiah yang diajukan dalam artikel ini didukung sepenuhnya oleh data yang benar-benar tersedia dan terverifikasi.



3.4 Praproses Data

Tahap praproses data (`src/01_cleaning.py`) menormalisasi teks ulasan dengan menghapus tautan (URL) dan derau (*noise*) yang tidak relevan, sembari mempertahankan penanda sosial yang bermakna seperti tagar (*hashtag*), penyebutan (*mention*), dan emoji, karena elemen-elemen ini terbukti membawa muatan sentimen pada teks media sosial informal. Skema data distandardisasi dan hasil praproses disimpan sebagai *output/clean_dataset.csv* beserta laporan audit *output/ig_01_audit.json* yang mendokumentasikan setiap transformasi yang diterapkan.

3.5 Rekayasa Fitur (*Feature Engineering*)

Rekayasa fitur (`src/04_feature_engineering.py`) menghasilkan representasi numerik teks melalui dua pendekatan yang digabungkan. Pertama, fitur berbasis leksikon dan sintaksis, mencakup panjang karakter, frekuensi tanda baca, rasio huruf kapital, keberagaman *token* (*token diversity*), serta skor leksikon sentimen bahasa Indonesia. Kedua, representasi TF-IDF (*Term Frequency-Inverse Document Frequency*) dengan rentang *n*-gram unigram dan bigram (1,2), yang menghasilkan total 5.130 fitur setelah digabungkan dengan fitur leksikal. Representasi gabungan ini dipilih untuk menangkap baik pola frekuensi kata sederhana maupun karakteristik gaya penulisan yang berpotensi berkorelasi dengan polaritas sentimen.

3.6 Skema Pembagian Data (*Data Splitting*)

Tahap pembagian data (`src/05_temporal_split.py`) awalnya dirancang untuk menerapkan skema pembagian temporal berdasarkan urutan waktu ulasan. Karena kolom cap waktu kontinu tidak tersedia pada dataset sebagaimana dijelaskan pada Bagian 3.3, tahap ini secara otomatis mengaudit ketersediaan kolom tanggal dan, ketika tidak ditemukan, beralih ke skema cadangan (*fallback*) berupa pembagian stratifikasi acak dengan rasio 80:20 antara data latih dan data uji, yang menjaga proporsi setiap kelas *rating* tetap konsisten antara kedua subset.

3.7 Arsitektur dan Konfigurasi Model

Empat algoritme klasifikasi dilatih dan dibandingkan pada representasi fitur yang identik: (1) Regresi Logistik sebagai *baseline* model linear; (2) *Random Forest* sebagai *ensemble bagging*; (3) XGBoost sebagai *ensemble boosting* dengan regularisasi eksplisit; dan (4) LightGBM sebagai *ensemble boosting* dengan strategi pertumbuhan pohon berbasis daun. Sebagai perbandingan tambahan, model Regresi Logistik juga dilatih di atas representasi TF-IDF murni (tanpa fitur leksikal tambahan) sebagai *baseline* sekunder. Artefak model terlatih disimpan dalam format *joblib* untuk memastikan reproduktibilitas eksperimen (`best_model.joblib`, `logistic_regression.joblib`, `random_forest.joblib`, `xgboost_model.joblib`, `lightgbm_model.joblib`, `tfidf_logistic_regression.joblib`).

3.8 Metrik Evaluasi

Evaluasi utama pada eksperimen yang tersedia menggunakan *accuracy* dan *weighted F1* pada hold-out test set. Karena distribusi kelas tidak seimbang, *accuracy* tidak digunakan sebagai satu-satunya dasar interpretasi. *Weighted F1* dipertahankan untuk merefleksikan kinerja agregat berbobot kelas, sedangkan macro-F1, *precision*, *recall*, F1 per kelas, dan *confusion matrix* ditetapkan sebagai evaluasi yang diperlukan untuk versi eksperimen berikutnya. Tidak ada nilai metrik yang belum dihitung yang direkonstruksi dari *accuracy* atau *weighted F1*.

Performa model dievaluasi pada data uji menggunakan *accuracy* dan *weighted F1*-score. Karena distribusi kelas tidak seimbang, macro-F1, *precision*, *recall*, F1 per kelas, dan *confusion matrix* seharusnya dilaporkan sebagai evaluasi lengkap. Pada versi eksperimen yang tersedia saat ini, hanya *accuracy* dan *weighted F1* yang terdokumentasi; metrik tambahan tidak diimputasi atau direkonstruksi dari angka agregat.

3.9 Kerangka Interpretabilitas: SHAP TreeExplainer

Interpretabilitas dianalisis menggunakan SHAP TreeExplainer untuk model berbasis pohon yang kompatibel. Analisis dilakukan pada dua tingkat: (1) *global feature importance*, yang merangkum kontribusi absolut rata-rata fitur terhadap prediksi; dan (2) *local explanation*, yang menjelaskan kontribusi fitur pada satu observasi. Regresi Logistik tidak diinterpretasikan



dengan TreeExplainer; jika diperlukan, metode *explainer* yang sesuai untuk model linear harus digunakan secara terpisah.

3.10 Interface Neural IndoBERT (Status: MISSING)

Sebagai bagian dari desain arsitektur penelitian, sebuah antarmuka (*interface*) untuk model bahasa IndoBERT (indobenchmark/indobert-base-p1) disiapkan (src/03_indobert.py) untuk memungkinkan ekstraksi fitur kontekstual neural sebagai pelengkap representasi TF-IDF pada penelitian lanjutan. Pada tahap penelitian ini, bobot model (*model weights*) belum ditempatkan secara lokal, sehingga status komponen ini dilaporkan secara transparan sebagai MISSING tanpa memicu pengunduhan otomatis berukuran besar yang di luar cakupan eksperimen saat ini. Komponen ini diposisikan sebagai fondasi untuk pengembangan penelitian lanjutan sebagaimana diuraikan pada Bagian 6.

3.11 Reproductibilitas dan Ketersediaan Kode

Seluruh kode sumber, dataset, model terlatih, dan hasil analisis SHAP pada penelitian ini tersedia secara terbuka pada repositori GitHub untuk mendukung reproductibilitas penelitian, sejalan dengan praktik ilmu data terbuka (*open science*) yang semakin menjadi standar dalam publikasi pada jurnal bereputasi. Setiap tahap *pipeline* dapat dijalankan secara independen maupun berurutan melalui skrip Python yang telah disediakan, memungkinkan peneliti lain memverifikasi atau memperluas temuan penelitian ini.

3.12 Pertimbangan Etis dan Privasi Data

Data ulasan yang digunakan dalam penelitian ini bersifat publik dan tidak memuat informasi identitas pribadi yang sensitif di luar nama pengguna (*username*) yang ditampilkan secara publik pada platform asal data. Sebagai langkah kehati-hatian tambahan, seluruh proses analisis dan pelaporan hasil dilakukan pada tingkat agregat dan *token* linguistik, tanpa menampilkan kutipan ulasan individual secara verbatim yang berpotensi mengidentifikasi pengguna tertentu secara langsung. Penulis menegaskan bahwa penelitian ini tidak melibatkan pengumpulan data baru dari subjek manusia secara langsung (*no direct human-subject data collection*), melainkan analisis sekunder atas data ulasan yang telah tersedia secara publik, sehingga risiko etis penelitian tergolong minimal. Meskipun demikian, penulis menyarankan konsultasi dengan komite etik penelitian institusi (jika dipersyaratkan oleh jurnal tujuan) sebelum naskah final diajukan, mengingat variasi kebijakan etik antar jurnal dan antar institusi.

3.13 Jaminan Kualitas Data (Data Quality Assurance)

Untuk memastikan kualitas data yang digunakan dalam eksperimen, penelitian ini menerapkan tiga lapisan pemeriksaan pada tahap pra proses: (1) pemeriksaan integritas skema, memastikan seluruh baris data memenuhi format kolom yang telah ditetapkan (*Username*, *Review Text*, *Rating*); (2) pemeriksaan kelengkapan, memverifikasi tidak adanya nilai kosong yang dapat mengganggu proses pelatihan model; dan (3) pemeriksaan duplikasi, mengidentifikasi dan menghapus baris duplikat pada kolom teks ulasan yang berpotensi menimbulkan kebocoran data (*data leakage*) antara subset data latih dan data uji apabila tidak ditangani. Ketiga lapisan pemeriksaan ini menghasilkan laporan audit terstruktur dalam format JSON yang disertakan sebagai bagian dari artefak reproductibilitas penelitian, sebagaimana dirinci pada Lampiran B.

3.14 Lingkungan Komputasi dan Perangkat Lunak

Seluruh eksperimen dijalankan menggunakan bahasa pemrograman Python versi 3.10 ke atas, dengan pustaka utama mencakup *scikit-learn* untuk model Regresi Logistik dan *Random Forest*, *xgboost* untuk model XGBoost, *lightgbm* untuk model LightGBM, serta *shap* untuk analisis interpretabilitas TreeExplainer. Dasbor penelitian interaktif dibangun menggunakan kerangka kerja Streamlit dengan pustaka visualisasi Plotly. Seluruh dependensi perangkat lunak didokumentasikan secara eksplisit pada berkas *requirements.txt* untuk menjamin reproductibilitas lingkungan komputasi oleh peneliti lain yang hendak memverifikasi atau memperluas temuan penelitian ini.



4. HASIL PENELITIAN

4.1 Performa Komparatif Model Klasifikasi

Tabel 3 menyajikan hasil evaluasi keempat model pada satu *hold-out test set*. *Accuracy* berada pada rentang 49,5-50,5% dan *weighted F1* pada rentang 41,5-48,3%. Regresi Logistik memiliki *weighted F1* tertinggi (48,3%), sedangkan Regresi Logistik dan XGBoost memiliki *accuracy* tertinggi yang sama (50,5%). Karena hanya satu split digunakan, perbedaan tersebut harus diperlakukan sebagai hasil empiris pada split yang tersedia, bukan sebagai bukti keunggulan model yang stabil.

Tabel 3. Perbandingan Performa Model pada Data Uji

Model	Akurasi	F1-Score Terboboti
Regresi Logistik	50,5%	48,3%
<i>Random Forest</i>	50,0%	43,1%
XGBoost	50,5%	41,5%
LightGBM	49,5%	41,5%

Temuan ini penting untuk digarisbawahi: pada dataset berukuran relatif kecil (1.000 sampel) dengan distribusi kelas yang timpang, kompleksitas model tambahan yang ditawarkan oleh algoritme *ensemble* tidak serta-merta menghasilkan peningkatan performa yang proporsional dibandingkan model linear yang jauh lebih sederhana secara komputasional. Diskusi lebih mendalam mengenai temuan ini disajikan pada Bagian 5.1.

4.2 Analisis Topik dan Distribusi Semantik

Analisis topik (`src/08_trend_momentum.py`) mengidentifikasi dua kluster topik dominan dalam korpus ulasan, dengan Topik 0 mencakup 54,9% ulasan dan Topik 1 mencakup 45,1% sisanya. Distribusi yang relatif seimbang antara kedua topik ini mengindikasikan bahwa korpus ulasan mencakup setidaknya dua tema besar keluhan atau apresiasi pengguna yang cukup berbeda secara semantik, meskipun keterbatasan ukuran sampel membatasi generalisasi lebih lanjut mengenai granularitas topik yang lebih spesifik.

4.3 Analisis Interpretabilitas Global: *SHAP Feature Importance*

Analisis SHAP TreeExplainer terhadap representasi fitur dilakukan untuk model berbasis pohon yang kompatibel. *Mean absolute SHAP* digunakan untuk merangkul kepentingan fitur secara global. Lima fitur teratas dilaporkan sebagai fitur dengan kontribusi absolut rata-rata terbesar; nilai tersebut tidak boleh ditafsirkan sebagai hubungan kausal atau sebagai bukti bahwa *token* tertentu secara linguistik menyebabkan *rating*.

Tabel 4. Lima Fitur Teratas Berdasarkan Nilai SHAP Global

Peringkat	<i>Token</i>	Nilai SHAP Rata-Rata
1	kenapa	0,103794
2	yang	0,086908
3	bisa	0,085416
4	bagus	0,078869
5	untuk	0,077572



Temuan ini menarik untuk dicermati karena empat dari lima *token* teratas ("kenapa", "yang", "bisa", "untuk") merupakan kata tanya dan kata penghubung yang secara linguistik netral, bukan kata bermuatan sentimen eksplisit seperti pada leksikon sentimen konvensional. Hanya *token* "bagus" yang secara intuitif membawa muatan sentimen positif yang jelas. Pola ini mengindikasikan bahwa model, khususnya varian berbasis pohon, kemungkinan besar memanfaatkan pola struktural kalimat keluhan (misalnya pertanyaan retorik yang mengindikasikan kekecewaan, seperti "kenapa aplikasi ini...") sebagai sinyal prediktif yang kuat, alih-alih murni bergantung pada leksikon sentimen eksplisit. Interpretasi ini dibahas lebih rinci pada Bagian 5.3.

4.4 Analisis Interpretabilitas Lokal: Studi Kasus Sampel

Selain analisis global, SHAP TreeExplainer juga menghasilkan penjelasan pada tingkat sampel individual (*local explanations*), yang memungkinkan penelusuran kontribusi *token* spesifik terhadap prediksi satu ulasan tertentu. Pendekatan ini memberikan nilai tambah praktis signifikan: pengembang aplikasi maupun analis dapat menelusuri secara spesifik mengapa sebuah ulasan diklasifikasikan sebagai negatif, alih-alih hanya menerima label akhir tanpa penjelasan. Kemampuan audit pada level sampel individual inilah yang membedakan pendekatan *explainable* dari pendekatan *black-box* konvensional, dan menjadi salah satu kontribusi utama yang ditawarkan penelitian ini bagi praktik industri.

4.5 Dasbor Riset Interaktif

Sebagai pelengkap laporan statis, penelitian ini juga mengembangkan dasbor penelitian interaktif berbasis Streamlit yang mengintegrasikan tiga belas bagian analisis: ringkasan umum (*overview*), audit dataset, antarmuka NLP/IndoBERT, intelijen topik, intelijen viralitas, tren dan momentum, *benchmark* pembelajaran mesin, eksplainabilitas SHAP, analisis jaringan, peramalan 2027, *pipeline* penelitian, sepuluh arah penelitian strategis, dan dokumentasi. Dasbor ini dirancang menggunakan kerangka visualisasi Plotly dengan kartu penelitian yang dapat diperluas (*expandable research cards*), pemeriksa SHAP sampel secara waktu nyata, serta lencana status (*status badges*) yang secara visual membedakan komponen berstatus AVAILABLE, PARTIAL, dan MISSING memungkinkan pengguna dasbor memahami tingkat keandalan setiap komponen analisis secara sekilas.

5. DISKUSI

5.1 Mengapa Model Kompleks Tidak Selalu Unggul

Hasil pada Bagian 4.1 menunjukkan bahwa Regresi Logistik memperoleh *weighted F1* tertinggi pada *split* yang digunakan. Salah satu penjelasan metodologis yang masuk akal adalah kecocokan model linear terregularisasi dengan representasi TF-IDF berdimensi tinggi; namun, penjelasan ini tetap bersifat interpretatif dan belum dapat diuji sebagai mekanisme kausal tanpa eksperimen robustness seperti *repeated cross-validation*, tuning yang terdokumentasi, dan analisis ablation.

5.2 Dampak Ketidakseimbangan Kelas terhadap Validitas Model

Distribusi kelas yang sangat timpang, dengan *rating* 1 mendominasi hampir separuh dataset, menjadi faktor pembatas utama performa keempat model yang diuji. Ketimpangan semacam ini mendorong model untuk cenderung memprediksi kelas mayoritas guna meminimalkan kesalahan agregat, sebuah fenomena yang dikenal luas dalam literatur pembelajaran mesin sebagai *majority class bias*. Akurasi 50,5% yang tampak moderat pada Tabel 3 perlu dibaca secara kritis: model yang secara naif selalu memprediksi kelas mayoritas (*rating* 1) tanpa mempelajari pola apa pun sekalipun akan mencapai akurasi dasar (*baseline*) sebesar 48,1%, sangat mendekati performa aktual keempat model yang diuji. Hal ini mengindikasikan bahwa kapasitas prediktif riil dari model-model tersebut, di luar sekadar mengikuti distribusi kelas mayoritas, kemungkinan besar jauh lebih terbatas daripada yang tersirat dari angka akurasi semata — sebuah temuan yang menegaskan pentingnya pelaporan



F1-score per kelas dan matriks konfusi pada revisi penelitian mendatang, sebagaimana diuraikan pada Bagian 6.

5.3 Interpretasi Linguistik atas Temuan SHAP

Lima fitur global teratas mencakup *token* yang tidak semuanya memiliki muatan sentimen eksplisit. Hal ini dapat mencerminkan kombinasi frekuensi *token*, konteks kemunculan, dan struktur representasi TF-IDF. Karena analisis saat ini tidak menyertakan *ablation* terhadap *stop-word handling* atau analisis *part-of-speech*, interpretasi retorik maupun linguistik tidak diperlakukan sebagai kesimpulan definitif.

5.4 Implikasi Praktis bagi Pengembang Aplikasi dan Peneliti UX

Terlepas dari keterbatasan performa akurasi, kontribusi interpretabilitas yang ditawarkan penelitian ini tetap memiliki nilai praktis yang signifikan. Kemampuan menelusuri kontribusi *token* pada level sampel individual (Bagian 4.4) memungkinkan tim produk dan UX *researcher* untuk tidak hanya mengetahui bahwa sebuah ulasan bersentimen negatif, tetapi juga memahami elemen bahasa spesifik apa yang memicu klasifikasi tersebut informasi yang jauh lebih *actionable* dibandingkan skor sentimen agregat semata. Kerangka audit transparansi status data (*AVAILABLE/PARTIAL/MISSING*) yang diterapkan sepanjang *pipeline* penelitian ini juga menawarkan model praktik baik (*best practice*) bagi tim *data science* industri yang kerap menghadapi tekanan untuk melaporkan hasil analitik tanpa mendokumentasikan keterbatasan data yang mendasarinya secara memadai.

5.5 Perbandingan dengan Studi Terdahulu

Studi	Data dan label	Pendekatan	Evaluasi/hasil utama	Apa yang berbeda dari studi ini
Penelitian ini	1.000 Instagram <i>reviews; rating</i> sebagai <i>proxy</i> label	TF-IDF + fitur leksikal; LR, RF, XGBoost, LightGBM; SHAP pada model tree	<i>Single stratified</i> 80:20; <i>accuracy</i> 49,5-50,5%; <i>weighted F1</i> 41,5-48,3%	Dataset kecil; <i>rating</i> -based label; belum k-fold
Aqilah et al. (2026)	7.187 Gojek <i>reviews; 14.187 sentence instances; manual 4-class annotation</i>	<i>IndoBERTweet; Optuna; sentence segmentation</i>	<i>Stratified k-fold; accuracy</i> 0,751; <i>macro-F1</i> 0,729	<i>Manual annotation</i> dan <i>contextual transformer</i> ; bukan <i>rating proxy</i>
Riccosan & Saputra (2025)	<i>Indonesian mobile app reviews; multilabel sentiment + emotion</i>	<i>IndoBERT; multilabel classification</i>	<i>Micro-F1</i> 0,77; ROC-AUC 0,85	<i>Multilabel sentiment-emotion</i> , bukan <i>tree ensemble + SHAP</i>
Fikri & Ridho (2025)	<i>Indonesian digital-media app reviews; review-rating inconsistency</i>	<i>Sentiment classification</i> untuk mengidentifikasi <i>inconsistency</i>	Fokus pada <i>mismatch review-rating</i>	Mendukung pentingnya memisahkan <i>rating</i> dari <i>textual sentiment</i>
Winanta et al. (2026)	<i>Indonesian e-commerce reviews</i>	<i>Random Forest + LIME/SHAP/ELI5 comparison</i>	<i>Comparative XAI analysis</i>	Lebih dekat dengan studi ini pada aspek XAI, tetapi domain dan eksperimen berbeda



Posisi penelitian ini karena itu lebih tepat dirumuskan sebagai studi audit dan *benchmarking explainable* berbasis *rating* pada korpus Instagram berbahasa Indonesia, bukan sebagai upaya untuk mengungguli *benchmark* IndoBERT/IndoBERTweet. *Novelty* yang dapat dipertahankan adalah integrasi *controlled comparison* empat *classifier* dengan audit SHAP dan pelaporan keterbatasan label/data secara eksplisit.

Perbandingan dengan studi terdahulu harus dibaca sebagai *contextual comparison*, bukan *head-to-head performance ranking*. Dataset, definisi label, jumlah kelas, *pre processing*, model, dan protokol validasi berbeda antar penelitian. Karena itu, angka performa berikut tidak digunakan untuk menyatakan bahwa satu penelitian atau model lebih unggul secara umum.

Penelitian	Data/Task	Model/Evaluasi	Hasil	Relevansi
Penelitian ini	1.000 <i>Instagram reviews; rating-based multiclass</i>	LR, RF, XGBoost, LightGBM; single stratified 80:20	<i>Accuracy</i> 49,5-50,5%; <i>weighted F1</i> 41,5-48,3%	Perlu <i>robustness test</i> dan metrik per kelas
Aqilah et al. (2026)	7.187 <i>Gojek reviews; sentence-level, 4 classes</i>	IndoBERTweet; stratified k-fold	<i>Accuracy</i> 0,751; <i>macro-F1</i> 0,729	<i>Cross-validation</i> dan desain anotasi lebih kuat
Riccosan & Saputra (2025)	<i>Indonesian mobile app reviews; multilabel sentiment + emotion</i>	IndoBERT; <i>multilabel evaluation</i>	<i>Micro-F1</i> 0,77; <i>ROC-AUC</i> 0,85	<i>Contextual representation</i> relevan untuk <i>app reviews</i>
Setyawati et al. (2026)	<i>Ruparupa reviews; weak rating labels</i>	<i>Rule-based aspect extraction + IndoBERT</i>	<i>Weighted F1</i> 0,61	Relevan untuk isu <i>weak/rating labeling</i>

Perbandingan berikut digunakan sebagai konteks literatur, bukan *direct head-to-head benchmark*, karena dataset, definisi label, jumlah kelas, arsitektur model, dan skema evaluasi berbeda antar studi. Aqilah et al. (2026) menggunakan 7.187 ulasan Gojek yang disegmentasi menjadi 14.187 unit kalimat dan melaporkan *accuracy* 0,751 serta *macro-F1* 0,729 melalui stratified k-fold. Riccosan dan Saputra (2025) menggunakan klasifikasi *multilabel sentiment-emotion* pada ulasan aplikasi dan melaporkan *micro-F1* 0,77 untuk IndoBERT. Setyawati et al. (2026) menggunakan *weak labeling* berbasis *rating* pada ulasan Ruparupa dan melaporkan *weighted F1* 0,61 untuk *sentiment classification*. Perbedaan tersebut menegaskan bahwa angka performa tidak dapat dibandingkan secara langsung dengan hasil penelitian ini; yang dapat dibandingkan adalah desain evaluasi, transparansi label, ukuran data, dan penggunaan interpretabilitas.

5.6 Ancaman terhadap Validitas Internal dan Eksternal

Sebagaimana lazim dalam penelitian eksperimental berbasis data, temuan penelitian ini perlu dibaca dengan mempertimbangkan sejumlah ancaman terhadap validitas. Dari sisi validitas internal, penggunaan skema pembagian data tunggal (*single train-test split*) tanpa validasi silang berlapis (*k-fold cross-validation*) berpotensi menghasilkan estimasi performa yang tidak stabil, mengingat ukuran data uji yang relatif kecil (sekitar 200 sampel pada rasio 80:20) rentan terhadap variasi acak dalam pembagian data. Dari sisi validitas eksternal, generalisasi temuan ke domain ulasan aplikasi lain di luar Instagram, atau ke platform media sosial yang berbeda karakteristik penggunaannya, perlu dilakukan dengan hati-hati mengingat korpus penelitian ini terbatas pada satu aplikasi dan satu titik waktu pengambilan data. Validitas konstruk juga perlu dicermati: skor *rating* bintang yang digunakan sebagai proksi label sentimen tidak selalu berkorelasi sempurna dengan polaritas sentimen tekstual yang



sesungguhnya diungkapkan dalam isi ulasan, sebuah fenomena *rating-text mismatch* yang telah didokumentasikan dalam literatur *app-review mining* sebelumnya.

5.7 Kontribusi terhadap Teori dan Praktik

Dari sisi konseptual, penelitian ini menunjukkan bahwa *explainability* tidak dapat dipisahkan dari kualitas label dan representasi fitur. SHAP dapat memperlihatkan fitur yang dominan dalam keputusan model, tetapi interpretasi tersebut tetap bergantung pada bagaimana label dibentuk dan bagaimana teks direpresentasikan. Dengan demikian, kontribusi penelitian ini lebih tepat diposisikan sebagai *methodological and empirical contribution* daripada pengembangan teori *sentiment analysis* baru.

5.8 Rekomendasi bagi Pengembang Aplikasi dan Praktisi Industri

Bagi tim pengembang aplikasi di Indonesia, temuan penelitian ini menyarankan beberapa langkah praktis. Pertama, sistem pemantauan sentimen otomatis idealnya tidak hanya melaporkan skor sentimen agregat, tetapi juga menyertakan penjelasan *token-level* sebagaimana dimungkinkan oleh SHAP, sehingga tim produk dapat menindaklanjuti keluhan secara lebih presisi. Kedua, mengingat temuan bahwa pola kalimat tanya retorik menjadi penanda kuat sentimen negatif, tim *customer support* dapat mengembangkan aturan penyaringan (*triage rules*) tambahan yang secara khusus menandai ulasan berpola tanya untuk peninjauan prioritas, tanpa perlu menunggu proses klasifikasi model pembelajaran mesin yang lebih kompleks selesai dilatih. Ketiga, mengingat keterbatasan performa model pada dataset kecil dan timpang, praktisi disarankan untuk tidak sepenuhnya menggantikan proses moderasi manual dengan sistem otomatis pada tahap kematangan model saat ini, melainkan memosisikan model sebagai alat bantu prioritisasi (*triage assistant*) yang mempercepat, bukan menggantikan, penilaian manusia.

7. KESIMPULAN

Penelitian ini mengevaluasi klasifikasi berbasis *rating* pada 1.000 ulasan aplikasi Instagram berbahasa Indonesia menggunakan Regresi Logistik, *Random Forest*, XGBoost, dan LightGBM dengan representasi TF-IDF dan fitur leksikal/sintaksis. Pada satu *stratified hold-out split* 80:20, *accuracy* berada pada rentang 49,5-50,5% dan *weighted F1* pada rentang 41,5-48,3%. Regresi Logistik menghasilkan *weighted F1* tertinggi pada split tersebut, tetapi hasil ini tidak ditafsirkan sebagai keunggulan umum karena belum terdapat *repeated/k-fold validation*. Analisis SHAP pada model *tree-based* mengidentifikasi lima fitur global dengan kontribusi absolut rata-rata terbesar, namun fitur tersebut tidak dapat ditafsirkan secara kausal. Kontribusi utama studi ini adalah *benchmarking* yang transparan dan auditabilitas interpretasi pada korpus *rating-based Indonesian app reviews*. Penelitian berikutnya perlu memvalidasi stabilitas hasil, menangani class imbalance, melaporkan metrik per kelas, dan membandingkan *rating-derived labels* dengan anotasi sentimen tekstual.

Melampaui kontribusi teknis dan empirisnya, penelitian ini menekankan pentingnya integritas dan transparansi data sebagai fondasi metodologis, dengan secara eksplisit mendokumentasikan keterbatasan dan kekosongan data alih-alih menyembunyikannya atau menggantinya dengan data sintetis. Prinsip ini tercermin dalam reorientasi fokus penelitian dari prediksi viralitas konten yang tidak dapat didukung secara valid oleh data yang tersedia menjadi klasifikasi sentimen ulasan aplikasi yang sepenuhnya berlandaskan bukti empiris yang dapat diverifikasi. Penulis berharap kerangka audit transparansi data dan pendekatan *explainable* yang diterapkan dalam penelitian ini dapat menjadi rujukan metodologis bagi peneliti lain yang bekerja dengan data media sosial dan ulasan pengguna berbahasa Indonesia, sekaligus mendorong praktik pelaporan ilmiah yang lebih jujur dan dapat direproduksi dalam bidang ilmu data sosial di Indonesia.



Secara keseluruhan, artikel ini disusun dengan semangat bahwa nilai ilmiah sebuah penelitian tidak semata-mata diukur dari ketinggian angka akurasi yang dilaporkan, melainkan juga dari kejujuran metodologis, kejelasan interpretasi, dan kesediaan peneliti untuk mendokumentasikan batas-batas pengetahuan yang dapat diklaim secara bertanggung jawab dari data yang tersedia. Prinsip inilah yang diharapkan dapat terus dipegang pada setiap tahap pengembangan naskah ini menuju bentuk akhirnya yang siap diajukan ke jurnal bereputasi.

UCAPAN TERIMA KASIH

Penulis menyampaikan terima kasih kepada Agung Budi Prasetyo (Institut Teknologi Telkom Surabaya), Hudan Studiawan (Institut Teknologi Sepuluh Nopember), dan Onno W. Purbo atas masukan dan bimbingan selama penyusunan naskah ini, serta kepada Program Studi Magister Ilmu Komunikasi UPN "Veteran" Jawa Timur atas dukungan yang diberikan.

DAFTAR PUSTAKA

- Breiman, L. (2001). *Random Forests*. *Machine learning*, 45(1), 5-32. DOI: 10.1023/A:1010933404324.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree *boosting* system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. DOI: 10.1145/2939672.2939785.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional *transformers* for language understanding. *Proceedings of NAACL-HLT 2019*, 4171-4186. DOI: 10.18653/v1/N19-1423.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient *boosting* decision tree. *Advances in Neural Information Processing Systems*, 30, 3146-3154.
- Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. *Proceedings of the 28th International Conference on Computational Linguistics*, 757-770. DOI: 10.18653/v1/2020.coling-main.66.
- Liu, B. (2012). *Sentiment analysis and opinion mining*. *Synthesis Lectures on Human Language Technologies*, 5(1), 1-167.
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765-4774.
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From *local explanations* to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56-67. DOI: 10.1038/s42256-019-0138-9.
- Pang, B., & Lee, L. (2008). Opinion mining and *sentiment analysis*. *Foundations and Trends in Information Retrieval*, 2(1-2), 1-135.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any *classifier*. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.