Jurnal Kolaborasi
Sains dan Ilmu Terapan
Volume 5 Nomor 1 Juli-Desember 2026 | Page 237-248
ISSN 2964- 5581 (media online)
https://utilityprojectsolution.org/ejournal/index.php/JuKSIT Penerbit : Cv. Utility Project Solutio

Perbandingan Model Deep Learning dan Transformer untuk Analisis Sentimen Ulasan Aplikasi Tokopedia

Restu Jati Saputro[1], Ratri Kurniasari[2], Hana Nurdina [3]
[1,3,] Program Studi D3 Administrasi Bisnis, Politeknik Negeri Jakarta, Indonesia
[2,] Program Studi D3 Administrasi Bisnis Terapan, Politeknik Negeri Jakarta, Indonesia

Email: [1] restujati.saputro@bisnis.pnj.ac.id, [2] ratri.kurniasari@bisnis.pnj.ac.id, [3,] hana.nurdina@bisnis.pnj.ac.id
Email Penulis Korespondensi: [1] restujati.saputro@bisnis.pnj.ac.id

Info Artikel

Kata Kunci:
Analisis sentimen
Tokopedia
weak label
IndoBERT
deep learning
Transformer

Abstrak

Rating bintang pada ulasan aplikasi tidak selalu merepresentasikan polaritas teks. Penelitian ini membandingkan tujuh model machine learning, deep learning, dan Transformer pada 10.071 ulasan Tokopedia berbahasa Indonesia dengan label teks berbasis InSet serta rating sebagai weak-label pembanding. InSet mencakup 9,447 ulasan; label akhir terdiri atas 5,342 negatif, 1,019 netral, dan 3,710 positif. Sebanyak 4,360 label InSet berbeda dari pemetaan rating. Pada lima seed, Linear SVM memperoleh mean macro-F1 tertinggi 0,711 ± 0,000, diikuti Logistic Regression 0,709 ± 0,000. Pada seed 42, bootstrap 95% CI model terbaik adalah [0,680, 0,742], sedangkan perbedaannya dengan runner-up tidak signifikan menurut McNemar-Holm (p=0,5666). Hasil menunjukkan bahwa kesimpulan benchmark sensitif terhadap skema label dan baseline linear tetap kompetitif.

Keywords:
Sentiment analysis;
Tokopedia;
weak labels;
IndoBERT;
deep learning;
Transformer.

Abstract

Star ratings in app reviews do not always represent textual polarity. This study compares seven machine-learning, deep-learning, and Transformer models on 10,071 Indonesian Tokopedia reviews using InSet-based textual labels and rating-derived weak labels as a comparator. InSet covered 9,447 reviews; the final labels comprised 5,342 negative, 1,019 neutral, and 3,710 positive instances. InSet disagreed with rating mapping on 4,360 reviews. Across five seeds, Linear SVM achieved the highest mean macro-F1 (0.711 ± 0.000), followed by Logistic Regression (0.709 ± 0.000). For seed 42, the best model's bootstrap 95% CI was [0.680, 0.742], while its difference from the runner-up was not significant under McNemar-Holm (p=0.5666). The benchmark conclusion is label-sensitive, and linear baselines remain competitive.

1. PENDAHULUAN

Ulasan aplikasi merupakan bentuk *user-generated content* yang merekam pengalaman pengguna secara langsung. Informasi di dalamnya dapat mengungkap berbagai persoalan, seperti bug, hambatan transaksi, kebutuhan fitur, serta persepsi terhadap kualitas layanan aplikasi. Tinjauan sistematis menunjukkan bahwa penambangan ulasan aplikasi telah banyak dimanfaatkan untuk mendukung aktivitas rekayasa perangkat lunak, terutama dalam identifikasi masalah, evaluasi fitur, dan perumusan kebutuhan pengguna [1]. Penelitian lain juga menekankan bahwa karakteristik ulasan perlu dipertimbangkan ketika menentukan strategi penambangan yang digunakan [2]. Dalam konteks e-commerce, klasifikasi sentimen pada ulasan produk menggunakan *deep learning* juga terbukti berguna untuk memahami persepsi pelanggan, meskipun objek ulasan produk berbeda dengan ulasan aplikasi [3]. Selain itu, tidak seluruh ulasan memiliki tingkat kebermanfaatan informasi yang sama sehingga proses penyaringan data menjadi bagian penting dalam analisis [4]. Oleh karena itu, penelitian ini secara khusus dibatasi pada ulasan aplikasi Tokopedia di Google Play, bukan pada ulasan produk atau toko di dalam marketplace.

Rating bintang menyediakan sumber label yang mudah diperoleh dalam jumlah besar, tetapi tidak selalu merepresentasikan polaritas yang terkandung dalam teks ulasan. Pengguna dapat memberikan rating rendah sambil tetap memuji aspek tertentu, memberikan rating tinggi meskipun menyampaikan keluhan, salah memilih jumlah bintang, atau membahas persoalan produk dan penjual yang tidak berkaitan langsung dengan aplikasi. Kajian mengenai *text-rating review discrepancy* menunjukkan bahwa rating dan isi teks dapat menyampaikan evaluasi yang berbeda [5]. Atas dasar itu, pemetaan rating 1–2 sebagai negatif, rating 3 sebagai netral, dan rating 4–5 sebagai positif dalam penelitian ini

Submitted: **29/05/2026**; Accepted: **20/06/2026**; Published:**22/07/2026**

Jurnal Kolaborasi
Sains dan Ilmu Terapan
Volume 5 Nomor 1 Juli-Desember 2026 | Page 237-248
ISSN 2964- 5581 (media online)
https://utilityprojectsolution.org/ejournal/index.php/JuKSIT Penerbit : Cv. Utility Project Solutio

diperlakukan sebagai *weak label*, bukan sebagai *ground truth*. Skema tersebut dapat digunakan untuk membentuk benchmark awal, tetapi validitasnya tetap perlu diperiksa melalui anotasi manusia.

Perkembangan analisis sentimen telah menghasilkan berbagai pendekatan dengan karakteristik dan kebutuhan komputasi yang berbeda. Model linear berbasis TF-IDF dapat digunakan sebagai baseline yang transparan, CNN mampu menangkap pola lokal berbentuk *n-gram*, sedangkan LSTM dan BiLSTM dirancang untuk memodelkan urutan dan dependensi antarkata. Di sisi lain, Transformer memanfaatkan representasi kontekstual yang telah dipelajari dari korpus besar. Berbagai penelitian pada ulasan konsumen, e-commerce, reputasi produk, dan *opinion mining* menunjukkan bahwa model *deep learning* dan Transformer dapat menghasilkan performa yang baik, tetapi hasilnya sangat dipengaruhi oleh karakteristik korpus, skema label, dan protokol evaluasi [6], [7]. Tinjauan terbaru juga mencatat sejumlah tantangan, seperti *domain shift*, data berisik, dan prosedur evaluasi yang tidak seragam [8], [9], [10]. Oleh sebab itu, keunggulan suatu arsitektur tidak dapat disimpulkan secara umum tanpa pengujian pada dataset dan kondisi eksperimen yang sama.

Analisis ulasan berbahasa Indonesia memiliki tantangan tambahan karena pengguna sering memadukan bahasa Indonesia dengan istilah Inggris, slang, singkatan, kesalahan ejaan, pengulangan karakter, emoji, serta konstruksi negasi. Kajian mengenai pengolahan bahasa alami di Indonesia menunjukkan tingginya keragaman bahasa dan dialek yang perlu diperhatikan dalam pengembangan model [11]. Sejumlah sumber daya telah dikembangkan untuk mendukung penelitian NLP Indonesia, di antaranya NusaCrowd dan NusaX yang menyediakan korpus dan benchmark untuk bahasa Indonesia serta bahasa daerah [11], [12]. IndoNLU dan IndoLEM juga menyediakan benchmark dan model bahasa pralatih untuk berbagai tugas pemrosesan bahasa Indonesia [13], [14]. Keberadaan sumber daya tersebut mendukung perbandingan model monolingual dan multilingual. IndoBERT dirancang khusus untuk bahasa Indonesia, sedangkan mBERT dan XLM-R berpotensi menangani ulasan yang mengandung campuran bahasa. Meskipun demikian, performa relatif ketiga model tetap harus dibuktikan melalui pengujian pada dataset yang sama.

Perbedaan arsitektur juga menjadi dasar penting untuk melakukan komparasi lintas keluarga model. CNN menggunakan filter konvolusional untuk mengekstraksi pola lokal yang diskriminatif, sedangkan LSTM mengatur aliran informasi melalui mekanisme memori dan gerbang [15]. BERT memperkenalkan encoder dua arah yang dapat diadaptasi untuk tugas klasifikasi teks [16]. mBERT memungkinkan transfer lintas bahasa melalui representasi multilingual, sedangkan XLM-R dikembangkan menggunakan data multibahasa dalam skala yang lebih besar [17]. RoBERTa kemudian menyempurnakan prosedur prapelatihan BERT melalui optimasi strategi pelatihan [18]. Perbedaan dalam representasi, tokenisasi, paralelisme, dan ukuran model tersebut dapat memengaruhi performa klasifikasi, ketahanan model, latensi inferensi, dan kebutuhan memori.

Sejumlah penelitian telah membahas analisis sentimen pada ulasan Tokopedia, tetapi ruang lingkup dan desain eksperimennya berbeda. BERT telah diterapkan pada ulasan aplikasi Tokopedia di Google Play [19]; sedangkan penelitian lain membandingkan pendekatan *machine learning* dan *word embedding* pada objek yang sama [20]. IndoBERT juga digunakan pada gabungan 12.000 ulasan aplikasi Lazada, Shopee, dan Tokopedia dengan weak label berbasis rating [21]. Selain itu, Naive Bayes dan KNN telah dibandingkan pada ulasan aplikasi e-commerce, termasuk Tokopedia [22], [23]. Studi lain membandingkan Random Forest dan LSTM pada dataset Tokopedia yang bersumber dari Kaggle [24]. Namun, penelitian tersebut menggunakan dua kelas dan lebih berorientasi pada ulasan produk atau layanan sehingga hasilnya tidak dapat dibandingkan secara langsung dengan ulasan aplikasi tiga kelas. Perbedaan sumber data, jumlah kelas, preprocessing, strategi pelabelan, dan pembagian data menyebabkan perbandingan skor antarpublikasi harus dilakukan secara hati-hati.

Literatur yang lebih luas juga menunjukkan pentingnya benchmark yang terkontrol. Arsitektur hibrida BERT-BiLSTM-CNN telah digunakan untuk klasifikasi ulasan aplikasi berbasis aspek [25]. sementara kombinasi BERT dengan CNN dan RNN telah diuji pada analisis sentimen e-commerce berbahasa Indonesia [26]. Perbandingan IndoBERT, mBERT, XLM-R, CNN, dan BiLSTM pada layanan pemerintah digital juga memberikan contoh evaluasi lintas keluarga model [27]. Sementara itu, penelitian pada media sosial Indonesia menunjukkan bahwa domain prapelatihan dan ukuran dataset dapat memengaruhi performa model IndoBERT pada teks informal [28]. Walaupun relevan, perbedaan objek, sumber data, dan tujuan klasifikasi menyebabkan penelitian-penelitian tersebut belum menjawab bagaimana tujuh model bekerja pada ulasan aplikasi Tokopedia dengan audit weak label, deduplikasi sebelum pembagian data, lima random seed, dan test set yang identik.

1.2 Penelitian terdahulu

Berdasarkan berbagai penelitian terdahulu, analisis sentimen pada ulasan aplikasi Tokopedia telah dilakukan menggunakan beragam pendekatan, mulai dari machine learning klasik hingga model Transformer. Namun, setiap penelitian memiliki perbedaan dalam sumber data, jumlah kelas, strategi pelabelan, cakupan model, dan protokol evaluasi. Perbedaan tersebut menyebabkan hasil antarpublikasi tidak dapat dibandingkan secara langsung. Oleh karena itu, diperlukan pemetaan yang lebih sistematis untuk menunjukkan posisi penelitian ini terhadap studi-studi sebelumnya. Ringkasan perbandingan penelitian terdekat disajikan pada Tabel 1

Submitted: **29/05/2026**; Accepted: **20/06/2026**; Published:**22/07/2026**

Jurnal Kolaborasi
Sains dan Ilmu Terapan
Volume 5 Nomor 1 Juli-Desember 2026 | Page 237-248
ISSN 2964- 5581 (media online)
https://utilityprojectsolution.org/ejournal/index.php/JuKSIT Penerbit : Cv. Utility Project Solutio

Tabel 1. Perbandingan studi terdekat dan posisi penelitian

Studi	Data/kelas	Model	Keterbatasan relatif
Setiawan dkk. (2024)	Tokopedia Google Play	BERT	Cakupan model dan audit label terbatas
Idris dkk. (2025)	Tokopedia app	ML + embedding	Belum mencakup semua Transformer
Aisy & Karyono (2025)	12.000; 3 app; 3 kelas; rating	IndoBERT	Lintas aplikasi; weak label
Tajuddin dkk. (2025)	Tokopedia + Shopee	Naive Bayes	Model tunggal
Wijianto dkk. (2025)	Tokopedia Google Play	KNN, NB	Tanpa neural/Transformer
Saputra & Budiman (2026)	Kaggle; 2 kelas	RF, LSTM	Domain/skema label berbeda
Marutho & Utomo (2025)	Layanan pemerintah	CNN, BiLSTM, IndoBERT, mBERT, XLM-R	Bukan Tokopedia
Penelitian ini	Tokopedia app; 10.071; 3 weak label	LR, SVM, CNN, BiLSTM, IndoBERT, mBERT, XLM-R	Tujuh model × lima seed untuk label rating dan InSet; override lexicon domain tetap terbuka

Tabel 1 menunjukkan bahwa penelitian terdahulu umumnya belum menggabungkan perbandingan lintas arsitektur, audit weak label, test set identik, dan evaluasi statistik dalam satu desain eksperimen. Oleh karena itu, gap penelitian ini terletak pada kebutuhan benchmark yang lebih terkontrol. Kebaruan penelitian ini bukan pada penggunaan Tokopedia atau Transformer, tetapi pada integrasi audit label rating dan InSet, deduplikasi sebelum split, lima random seed, bootstrap, McNemar-Holm, analisis kesalahan, dan pengukuran efisiensi komputasi. Desain ini juga mempertimbangkan bahwa satu ulasan dapat memuat lebih dari satu sinyal dan kosakata domain aplikasi [29], [30]. Penelitian ini bertujuan membandingkan performa model klasik, deep learning, dan Transformer pada ulasan Tokopedia berbahasa Indonesia. RQ1 menilai model terbaik, RQ2 mengukur pengaruh perbedaan label rating dan InSet, RQ3 menguji signifikansi perbedaan model, RQ4 mengidentifikasi pola kesalahan, dan RQ5 menilai trade-off performa dan efisiensi.

2. METODOLOGI PENELITIAN

2.1 Desain, sumber, dan etika data

Penelitian ini menggunakan desain kuantitatif eksperimental-komparatif dengan satu pipeline data dan pembagian dataset yang sama untuk seluruh model. Data berupa ulasan publik aplikasi Tokopedia di Google Play dengan package ID com.tokopedia.tkpd, bahasa dan negara id/ID, pada periode Januari–Juni 2026. Metadata aplikasi diverifikasi menggunakan google-play-scraper, kemudian ulasan diambil dengan Sort.NEWEST, batch 200, jeda satu detik, maksimal tiga percobaan ulang, dan checkpoint setiap 1.000 ulasan. Proses dihentikan setelah data melewati batas awal periode penelitian. Data mentah disimpan dalam format CSV UTF-8 BOM, JSONL, dan XLSX, sedangkan continuation token dan fingerprint konfigurasi digunakan untuk mendukung proses resume. Atribut yang disimpan meliputi review_id, review_text, rating, jumlah thumbs-up, waktu ulasan, balasan pengembang, versi aplikasi, sumber, dan waktu pengambilan. Nama pengguna hanya disimpan pada data mentah untuk keperluan audit dan dihapus dari dataset terproses serta sampel anotasi. Pengumpulan data tidak menggunakan login, bypass CAPTCHA, proxy berputar, atau mekanisme penghindaran pembatasan. Dataset diperlakukan sebagai snapshot hasil scraper pihak ketiga, bukan sebagai arsip lengkap Google Play.

2.2 Pembersihan, pelabelan InSet, dan deduplikasi

Pipeline mempertahankan text_raw untuk model Transformer dan membentuk text_clean untuk model linear, CNN, dan BiLSTM melalui normalisasi Unicode, *lowercasing*, pembersihan HTML/URL, serta normalisasi spasi. Emoji dan negasi tetap dipertahankan karena dapat membawa informasi sentimen. Duplikat identik dihapus sebelum pembagian data agar satu kelompok ulasan tidak muncul pada lebih dari satu subset. Rating 1–2, 3, dan 4–5 digunakan sebagai *weak label* pembanding. Label eksperimen utama dibentuk menggunakan InSet [31], yang memberikan bobot polaritas −5 hingga +5. Pencocokan dilakukan dari frasa terpanjang, sedangkan negasi dalam tiga token sebelumnya membalik polaritas dan *intensifier* meningkatkan kekuatan skor. Skor positif, negatif, dan nol masing-masing dipetakan menjadi label positif, negatif, dan netral. Workbook domain digunakan untuk menyesuaikan atau mengabaikan entri InSet tanpa mengubah lexicon asli.

Tabel 2. Perbedaan input teks antarkeluarga model

Tahap	text_raw (Transformer)	text_clean (linear/CNN/BiLSTM)
Case	Dipertahankan	Lowercase
URL/email/HTML	Dipertahankan sebagai teks asli	Dihapus
Spasi	Dipertahankan	Dinormalisasi
Simbol	Dipertahankan	Non-alfanumerik tak perlu dihapus
Emoji	Dipertahankan	Diekstrak ke emoji_text

Submitted: **29/05/2026**; Accepted: **20/06/2026**; Published:**22/07/2026**

Jurnal Kolaborasi
Sains dan Ilmu Terapan
Volume 5 Nomor 1 Juli-Desember 2026 | Page 237-248
ISSN 2964- 5581 (media online)
https://utilityprojectsolution.org/ejournal/index.php/JuKSIT Penerbit : Cv. Utility Project Solutio

Tahap	text_raw (Transformer)	text_clean (linear/CNN/BiLSTM)
Negasi	Dipertahankan	Dipertahankan

Tabel 2 menunjukkan bahwa Transformer menggunakan teks yang lebih mendekati bentuk asli agar informasi kontekstual tetap terjaga. Sebaliknya, model linear dan neural menerima teks yang telah dibersihkan untuk mengurangi noise dan menghasilkan representasi yang lebih konsisten. Meskipun demikian, emoji dan negasi tetap dipertahankan pada kedua jalur karena keduanya dapat memengaruhi polaritas sentimen.

2.3 Audit lexicon dan pembagian data

InSet menemukan sedikitnya satu kecocokan pada 9.447 dari 10.071 ulasan atau 93,80%. Sebanyak 624 ulasan tanpa kecocokan diberi skor nol dan dikategorikan sebagai netral. Label InSet berbeda dari weak label berbasis rating pada 4.360 ulasan atau 43,29%. Nilai tersebut menunjukkan ketidaksesuaian antarskema pelabelan, bukan tingkat kesalahan terhadap gold label manusia. Untuk kedua skema label, data dibagi menjadi 7.050 data latih, 1.511 data validasi, dan 1.510 data uji dengan seed 42. Seluruh model dan lima seed menggunakan test set yang sama. Data validasi digunakan untuk early stopping, sedangkan test set tidak digunakan dalam proses tuning.

2.4 Model, konfigurasi, dan evaluasi

Tujuh model diuji menggunakan test set yang sama. Logistic Regression dan Linear SVM menggunakan gabungan TF-IDF kata dan karakter, sedangkan CNN dan BiLSTM dilatih dari text_clean dengan panjang maksimum 64 token. Tiga Transformer menggunakan text_raw dengan encoder beku dan kuantisasi dinamis INT8 karena eksperimen dijalankan pada CPU. Oleh karena itu, hasil Transformer merepresentasikan *frozen-encoder transfer learning*, bukan *full fine-tuning*. Macro-F1 digunakan sebagai metrik utama, disertai accuracy, macro-precision, macro-recall, weighted-F1, dan F1 per kelas. Nilai rata-rata dan standar deviasi dihitung dari seed 42–46. Ketidakpastian hasil dievaluasi melalui 2.000 bootstrap terstratifikasi, sedangkan perbedaan antarmodel diuji menggunakan McNemar dengan koreksi Holm [32], [33], [34].

Analisis kesalahan mencakup negasi, emoji, pengulangan karakter, code-mixing, ulasan sangat pendek, teks panjang, dan kandidat mismatch. Efisiensi model diukur berdasarkan waktu pelatihan, latensi inferensi, peak RSS, jumlah parameter, dan ukuran bobot. Kalibrasi probabilitas dievaluasi secara terpisah karena akurasi tidak selalu mencerminkan tingkat kepercayaan prediksi [35].

Tabel 3. Model dan konfigurasi eksperimen

Model	Input	Konfigurasi yang dieksekusi/status
Logistic Regression	word+char TF-IDF; text_clean	C=4; balanced; liblinear; OvR
Linear SVM	word+char TF-IDF; text_clean	C=1; balanced; max_iter=5.000
CNN	token; text_clean	emb=128; 128 filter × {3,4,5}; dropout=0,5
BiLSTM	token; text_clean	emb=128; hidden=128 bidirectional; dropout=0,5
IndoBERT-Lite	subword; text_raw	encoder beku INT8; maxlen=64; head linear
mBERT	subword; text_raw	encoder beku INT8; maxlen=64; head linear
XLM-R	SentencePiece; text_raw	encoder beku INT8; maxlen=64; head linear

Tabel 3 merangkum perbedaan representasi input dan konfigurasi utama setiap model. Model linear menggunakan fitur TF-IDF, CNN dan BiLSTM mempelajari representasi token, sedangkan Transformer memanfaatkan embedding kontekstual dari encoder pralatih yang dibekukan

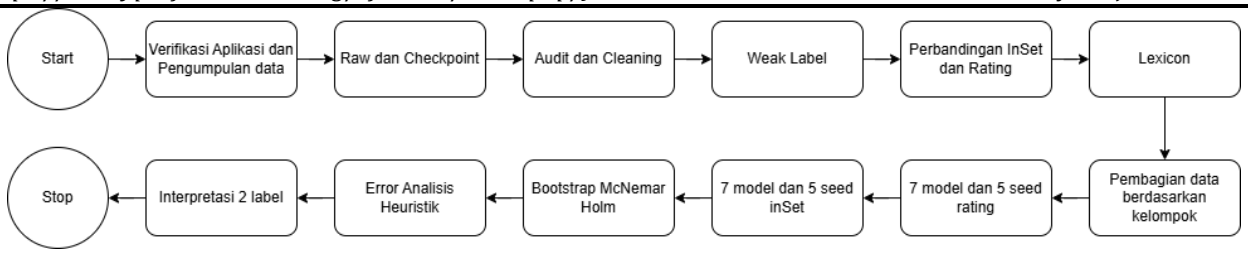
2.5 Perangkat dan proses pengolahan

Pipeline data dan eksperimen dengan seed 42–46 dijalankan pada Windows 11 Pro 64-bit menggunakan Intel Core i7-8665U, RAM 15,74 GB, dan Intel UHD Graphics 620 tanpa GPU diskrit. Eksperimen menggunakan Python 3.13.11, PyTorch 2.13.0, Transformers 5.13.1, scikit-learn 1.9.0, SciPy 1.18.0, dan SentencePiece 0.2.2. Untuk mendukung pengolahan proyek menyimpan kode, konfigurasi, daftar dependensi, unit test, pembagian data, file prediksi, log eksperimen, cache embedding, dan artefak model dilakukan.

.

Submitted: **29/05/2026**; Accepted: **20/06/2026**; Published:**22/07/2026**

Jurnal Kolaborasi
Sains dan Ilmu Terapan
Volume 5 Nomor 1 Juli-Desember 2026 | Page 237-248
ISSN 2964- 5581 (media online)
https://utilityprojectsolution.org/ejournal/index.php/JuKSIT Penerbit : Cv. Utility Project Solutio



Gambar 1. Diagram alur penelitian

Gambar 1 memperlihatkan tahapan penelitian mulai dari verifikasi aplikasi, pengumpulan dan pembersihan data, pembentukan weak label, perbandingan InSet dan rating, pembagian data, evaluasi tujuh model pada lima seed, pengujian statistik, analisis kesalahan, hingga interpretasi hasil kedua skema label

3. HASIL DAN PEMBAHASAN

3.1 Audit dataset dan distribusi label

Dari 12.593 ulasan mentah, sebanyak 10.071 ulasan dipertahankan setelah pembersihan dan penghapusan 2.434 duplikat. Dataset kemudian dibagi menjadi 7.050 data latih, 1.511 data validasi, dan 1.510 data uji tanpa kebocoran review_id maupun duplicate_group. Pelabelan berbasis rating menghasilkan 5.373 ulasan negatif, 624 netral, dan 4.074 positif. Sementara itu, InSet menghasilkan 5.342 negatif, 1.019 netral, dan 3.710 positif. Peningkatan kelas netral pada InSet terutama berasal dari skor nol dan ulasan tanpa kecocokan lexicon, sehingga evaluasi perlu mempertimbangkan macro-F1 dan F1 tiap kelas.

Tabel 4. Karakteristik dataset hasil pipeline

Indikator	Nilai
Raw review	12.593
Data bersih	10.071 (79,97%)
Anggota grup duplikat	2.707
Duplikat dihapus	2.434
Hanya simbol/noise	85
Teks bersih <3 karakter	298
Kandidat mismatch	45 (0,45% data bersih)
Panjang teks, mean/median	97,08 / 60 karakter
Balasan pengembang	430 (4,27%)
Informasi versi aplikasi	7.534 (74,81%)

Tabel 4 merangkum proses pembersihan data dan karakteristik utama dataset. Hasilnya menunjukkan bahwa sekitar 80% ulasan mentah dapat digunakan untuk analisis setelah penghapusan duplikat dan data tidak valid

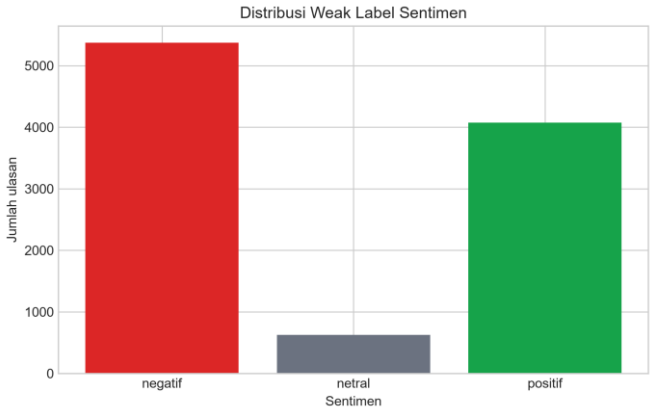
Tabel 5. Distribusi weak label dan split

Label	Rating	Jumlah	Persentase	Train	Validation	Test
Negatif	1–2	5.373	53,35%	3.761	806	806
Netral	3	624	6,20%	437	94	93
Positif	4–5	4.074	40,45%	2.852	611	611
Total	1–5	10.071	100,00%	7.050	1.511	1.510

Tabel 5 menunjukkan bahwa kelas negatif mendominasi dataset, sedangkan kelas netral memiliki proporsi paling kecil. Ketidakseimbangan ini menjadi alasan penggunaan macro-F1 sebagai metrik utama.

Submitted: **29/05/2026**; Accepted: **20/06/2026**; Published:**22/07/2026**

Jurnal Kolaborasi
Sains dan Ilmu Terapan
Volume 5 Nomor 1 Juli-Desember 2026 | Page 237-248
ISSN 2964- 5581 (media online)
https://utilityprojectsolution.org/ejournal/index.php/JuKSIT Penerbit : Cv. Utility Project Solutio



Gambar 2. Distribusi weak label pada data bersih

Gambar 2 memperlihatkan perbedaan jumlah ulasan pada kelas negatif, netral, dan positif. Distribusi yang tidak seimbang, terutama rendahnya jumlah kelas netral.

3.2 Karakteristik temporal, panjang teks, dan metadata

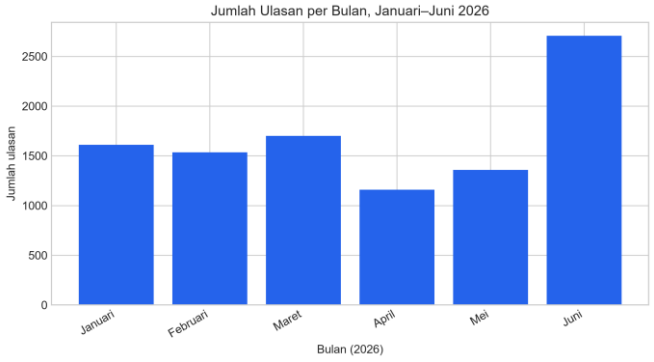
Data mencakup ulasan Januari–Juni 2026 sebagaimana dikembalikan scraper tanpa offset zona waktu per ulasan. Rata-rata panjang text_raw adalah 97,08 karakter dengan median 60 karakter, yang menunjukkan bahwa sebagian besar ulasan relatif singkat, sementara sejumlah ulasan panjang menaikkan nilai rata-rata. Ulasan pendek cenderung lebih ambigu, sedangkan ulasan panjang dapat memuat sentimen campuran. Karakteristik ini penting untuk menentukan panjang sekuens dan menganalisis kesalahan berdasarkan panjang teks.

Sebanyak 430 ulasan atau 4,27% memperoleh balasan pengembang, sedangkan 7.534 ulasan atau 74,81% memuat informasi versi aplikasi. Kedua atribut dapat digunakan untuk analisis lanjutan, tetapi ukuran sampel tiap kelompok perlu dilaporkan karena data yang hilang tidak selalu terjadi secara acak.

Tabel 6. Review bersih dan rata-rata rating per bulan

Bulan (2026)	Review bersih	Rata-rata rating
Januari	1610	2.871
Februari	1535	2.740
Maret	1701	2.708
April	1160	2.369
Mei	1359	2.396
Juni	2706	3.059

Tabel 6 menunjukkan bahwa jumlah ulasan tertinggi terjadi pada Juni, sedangkan rata-rata rating terendah tercatat pada April. Perubahan tersebut mengindikasikan adanya variasi pengalaman pengguna antarbulan.



Gambar 3. Jumlah ulasan bersih per bulan

Submitted: **29/05/2026**; Accepted: **20/06/2026**; Published:**22/07/2026**

Jurnal Kolaborasi
Sains dan Ilmu Terapan
Volume 5 Nomor 1 Juli-Desember 2026 | Page 237-248
ISSN 2964- 5581 (media online)
https://utilityprojectsolution.org/ejournal/index.php/JuKSIT Penerbit : Cv. Utility Project Solutio

Gambar 3 memperlihatkan fluktuasi jumlah ulasan selama Januari - Juni 2026. Peningkatan paling besar terjadi pada Juni, sementara April memiliki jumlah ulasan paling sedikit

3.4 Audit ketidaksesuaian rating - teks

Sebanyak 4.360 dari 10.071 ulasan atau 43,29% memiliki label InSet yang berbeda dari pemetaan rating. Nilai ini menunjukkan perbedaan hasil antara dua skema pelabelan, bukan kesalahan rating maupun kebenaran InSet. Karena InSet dikembangkan dari data microblog, penggunaannya pada ulasan e-commerce berpotensi mengalami *domain shift*. Oleh karena itu, hasil dilaporkan sebagai label berbasis lexicon, sedangkan workbook koreksi domain disediakan untuk meninjau istilah yang tidak sesuai konteks.

3.5 Performa, signifikansi, dan efisiensi model

Linear SVM memperoleh mean macro-F1 tertinggi sebesar 0,711 ± 0,000, diikuti Logistic Regression sebesar 0,709 ± 0,000. CNN dan BiLSTM masing-masing mencapai 0,680 ± 0,012, sedangkan IndoBERT-Lite, mBERT, dan XLM-R memperoleh nilai lebih rendah pada konfigurasi encoder beku.

Pada seed 42, Linear SVM menghasilkan macro-F1 0,711 dengan 95% CI bootstrap [0,680, 0,742]. Selisihnya dengan Logistic Regression tidak signifikan berdasarkan uji McNemar-Holm (p = 0,5666). Dengan demikian, kedua baseline linear memiliki performa yang relatif setara.

Confusion matrix menunjukkan bahwa Linear SVM mengenali 65 dari 136 data netral. Dari sisi efisiensi, Linear SVM juga memiliki waktu pelatihan tercepat dan ukuran artefak terkecil. Hasil Transformer perlu dibaca sebagai *frozen-encoder transfer learning*, bukan *full fine-tuning*, sehingga performanya tidak dapat digeneralisasi untuk seluruh konfigurasi Transformer.

Pergantian skema label dari rating ke InSet juga mengubah skor dan peringkat model. Hal ini menunjukkan bahwa definisi label berpengaruh terhadap hasil benchmark dan perlu diaudit secara eksplisit.

Tabel 7A. Performa model pada test set yang sama (mean ± SD seed 42–46)

Model	Accuracy mean±SD	P-macro mean±SD	R-macro mean±SD	F1-macro mean±SD	F1-weighted mean±SD	F1 Neg/Neu/Pos seed 42	95% CI F1 seed 42
Logistic Regression	0.786±0.000	0.707±0.000	0.712±0.000	0.709±0.000	0.787±0.000	0.840/0.511/0.778	[0.681, 0.739]
Linear SVM	0.790±0.000	0.720±0.000	0.703±0.000	0.711±0.000	0.788±0.000	0.838/0.510/0.785	[0.680, 0.742]
CNN	0.745±0.010	0.673±0.014	0.694±0.014	0.680±0.012	0.747±0.008	0.811/0.538/0.718	[0.660, 0.717]
BiLSTM	0.751±0.014	0.670±0.013	0.703±0.012	0.680±0.012	0.756±0.011	0.829/0.506/0.747	[0.666, 0.721]
IndoBERT-Lite	0.604±0.012	0.540±0.008	0.590±0.004	0.550±0.009	0.614±0.010	0.662/0.400/0.590	[0.522, 0.577]
mBERT	0.559±0.015	0.503±0.010	0.538±0.018	0.512±0.013	0.566±0.015	0.653/0.380/0.523	[0.493, 0.545]
XLM-R	0.514±0.018	0.464±0.013	0.507±0.010	0.465±0.021	0.524±0.017	0.651/0.308/0.357	[0.414, 0.463]

Tabel 7A menunjukkan bahwa Linear SVM dan Logistic Regression memberikan performa terbaik, sedangkan kelas netral tetap menjadi kelas yang paling sulit dikenali. Nilai interval kepercayaan diperoleh melalui 2.000 bootstrap terstratifikasi pada test set seed 42

Tabel 7B. Efisiensi komputasi model (seed 42, CPU)

Model	Fit	Inferensi b1	Inferensi b16	Peak RSS	Bobot	Parameter
Logistic Regression	5.71 s	4.61 ms	0.586 ms	459 MB	1.09 MB	0.116 M
Linear SVM	4.01 s	3.05 ms	0.474 ms	472 MB	1.02 MB	0.116 M
CNN	69.62 s	2.16 ms	0.570 ms	474 MB	5.20 MB	1.364 M
BiLSTM	110.00 s	3.51 ms	0.905 ms	483 MB	5.46 MB	1.431 M
IndoBERT-Lite	2180.00 s	503.64 ms	187.762 ms	443 MB	44.58 MB	11.686 M
mBERT	659.34 s	143.36 ms	65.514 ms	443 MB	678.47 MB	177.856 M
XLM-R	670.38 s	153.22 ms	59.459 ms	443 MB	1060.66 MB	278.046 M

Tabel 7B menunjukkan bahwa Linear SVM merupakan model paling efisien dari sisi waktu pelatihan dan ukuran bobot. Sebaliknya, Transformer membutuhkan latensi dan penyimpanan yang jauh lebih besar meskipun encoder dibekukan dan dikuantisasi INT8.

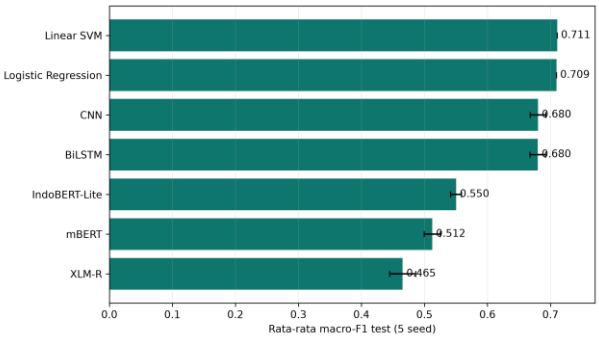
Tabel 8. Hasil uji statistik dan analisis lanjutan

Analisis	Output	Hasil aktual
Multi-seed	Mean ± SD, 5 seed	Terbaik: Linear SVM 0.711 ± 0.000
Bootstrap 2.000×	95% CI macro-F1 Linear SVM	[0.680; 0.742]
Selisih bootstrap	95% CI Linear SVM–Logistic Regression	[-0.0136; 0.0164]
McNemar + Holm	Linear SVM vs enam model	5/6 signifikan; vs Logistic Regression p=0.5666
Kalibrasi	ECE terendah	IndoBERT-Lite 0.024
Error heuristik	Tertinggi n≥20	code_mixing_sederhana: 24.6% (48/195)

Tabel 8 menegaskan bahwa keunggulan Linear SVM atas Logistic Regression tidak signifikan. IndoBERT-Lite menunjukkan kalibrasi terbaik, sedangkan code-mixing menjadi kategori kesalahan heuristik tertinggi

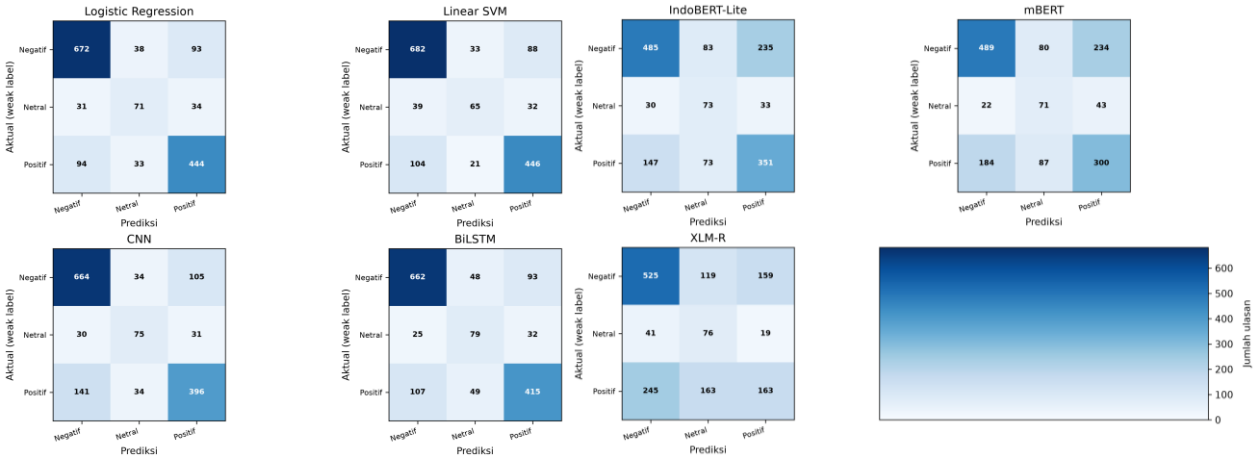
Submitted: **29/05/2026**; Accepted: **20/06/2026**; Published:**22/07/2026**

Jurnal Kolaborasi
Sains dan Ilmu Terapan
Volume 5 Nomor 1 Juli-Desember 2026 | Page 237-248
ISSN 2964- 5581 (media online)
https://utilityprojectsolution.org/ejournal/index.php/JuKSIT Penerbit : Cv. Utility Project Solutio



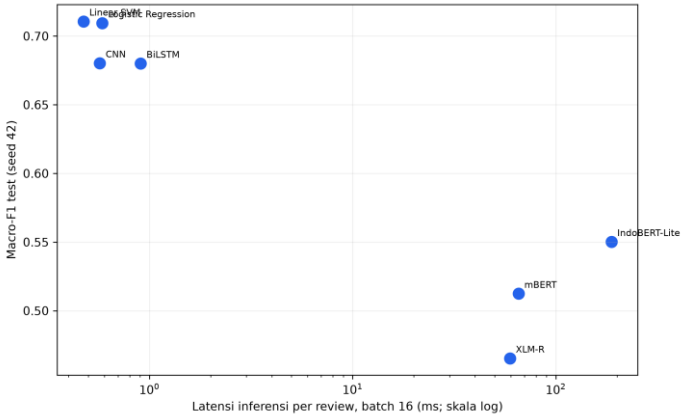
Gambar 4. Perbandingan macro-F1 tujuh model yang selesai dievaluasi

Gambar 4 menunjukkan bahwa Linear SVM dan Logistic Regression memiliki macro-F1 tertinggi, sedangkan ketiga Transformer beku berada di bawah model linear dan neural.



Gambar 5. Confusion matrix tujuh model terhadap label InSet

Gambar 5 memperlihatkan pola kesalahan tiap model pada kelas negatif, netral, dan positif. Kelas netral menjadi kelas yang paling sulit dikenali oleh sebagian besar model.

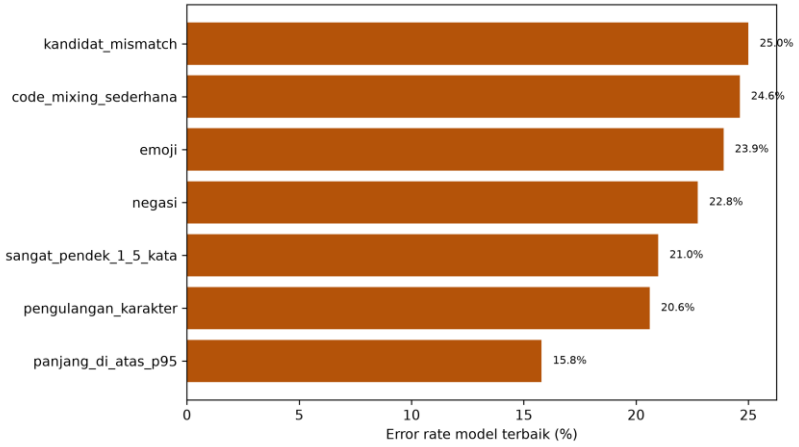


Gambar 6. Trade-off macro-F1 dan latensi inferensi batch 16

Gambar 6 menunjukkan bahwa model linear memberikan keseimbangan terbaik antara performa dan kecepatan inferensi. Transformer memiliki latensi lebih tinggi dengan macro-F1 yang lebih rendah pada konfigurasi encoder beku.

Submitted: **29/05/2026**; Accepted: **20/06/2026**; Published:**22/07/2026**

Jurnal Kolaborasi
Sains dan Ilmu Terapan
Volume 5 Nomor 1 Juli-Desember 2026 | Page 237-248
ISSN 2964- 5581 (media online)
https://utilityprojectsolution.org/ejournal/index.php/JuKSIT Penerbit : Cv. Utility Project Solutio



Gambar 7. Error rate subset heuristik pada Linear SVM

Gambar 7 memperlihatkan bahwa *code - mixing, mismatch rating* - teks, emoji, dan negasi merupakan sumber kesalahan utama pada Linear SVM

Model linear menunjukkan performa yang kompetitif, tetapi skor antarpublikasi tidak dapat dibandingkan langsung karena perbedaan korpus, preprocessing, dan skema label. Penggunaan InSet pada ulasan aplikasi juga berpotensi mengalami *domain shift* karena lexicon tersebut dikembangkan dari data microblog. Pada dua skema label, Transformer dengan encoder beku belum melampaui model linear. Hasil ini hanya berlaku pada konfigurasi CPU, panjang maksimum 64 token, kuantisasi INT8, dan tanpa *full fine-tuning*. Oleh karena itu, temuan tersebut tidak dapat digeneralisasi sebagai bukti bahwa model linear selalu lebih baik daripada Transformer.

Perubahan skema label memengaruhi skor dan peringkat model. Macro-F1 Logistic Regression meningkat dari 0,617 pada label rating menjadi 0,709 pada InSet, sedangkan Linear SVM meningkat dari 0,596 menjadi 0,711. CNN dan BiLSTM juga meningkat, tetapi IndoBERT-Lite dan XLM-R mengalami penurunan. Temuan ini menunjukkan bahwa fungsi label mengubah batas keputusan dan contoh yang harus dipelajari model. Keunggulan model linear kemungkinan dipengaruhi oleh kesesuaian antara fitur TF-IDF dan mekanisme InSet yang sama-sama berbasis kata atau frasa. Namun, hasil tersebut belum membuktikan bahwa SVM lebih baik dalam memahami sentimen alami. Evaluasi terhadap anotasi manusia tetap diperlukan.

CNN dan BiLSTM memperoleh macro-F1 yang sama, yaitu 0,680 ± 0,012. Perbedaan yang kecil dan simpangan baku yang bertumpang tindih tidak mendukung klaim bahwa salah satu model secara konsisten lebih unggul. Sementara itu, IndoBERT-Lite, mBERT, dan XLM-R hanya menguji transfer fitur dari encoder beku, bukan kemampuan *full fine-tuning* [13], [14], [16], [17], [36]. Kelas netral juga perlu ditafsirkan dengan hati-hati karena skor InSet nol dapat menunjukkan teks netral, sentimen campuran, atau tidak adanya kata yang tercakup lexicon. Oleh karena itu, peningkatan F1 kelas netral belum tentu menunjukkan pemahaman sentimen yang lebih baik.

Uji McNemar-Holm menunjukkan bahwa perbedaan Linear SVM dan Logistic Regression tidak signifikan (p = 0,5666). Interval bootstrap 95% Linear SVM sebesar [0,680; 0,742] juga menunjukkan adanya ketidakpastian pada sampel uji. Pelaporan mean ± SD dan bootstrap memberikan informasi yang saling melengkapi mengenai variasi seed dan sampel

[33], [34]. Dibandingkan studi Tokopedia sebelumnya [19], [20], [21], [22], [23] , kontribusi utama penelitian ini bukan sekadar skor yang lebih tinggi, tetapi bukti bahwa definisi label dapat mengubah hasil benchmark. Karena itu, sumber label harus diperlakukan sebagai bagian penting dari desain eksperimen.

3.6 Implikasi dan keterbatasan

Secara praktis, Logistic Regression dan Linear SVM dapat digunakan untuk penyaringan awal karena memiliki performa, ukuran model, dan latensi yang efisien. Namun, prediksi dengan confidence rendah, ulasan tanpa kecocokan InSet, serta kasus ketidaksesuaian rating dan teks tetap perlu diperiksa manusia. Kontribusi penelitian ini mencakup dua skema label yang dapat diaudit, split identik, lima seed, 70 run, bootstrap, McNemar-Holm, analisis kesalahan, dan evaluasi efisiensi komputasi.

Keterbatasan utama terletak pada penggunaan InSet sebagai label berbasis lexicon, bukan gold label manusia. Selain itu, aturan negasi, intensifier, dan penanganan ironi masih sederhana. Ulasan kontras seperti “bagus tetapi pembayaran gagal” juga dapat dinilai berdasarkan jumlah kata, bukan maksud utama pengguna.

Validitas eksternal dibatasi oleh penggunaan satu aplikasi dan periode Januari–Juni 2026. Perubahan fitur, promosi, atau gangguan layanan dapat mengubah kosakata dan distribusi kelas. Oleh karena itu, model perlu diuji ulang

Submitted: **29/05/2026**; Accepted: **20/06/2026**; Published:**22/07/2026**

Jurnal Kolaborasi
Sains dan Ilmu Terapan
Volume 5 Nomor 1 Juli-Desember 2026 | Page 237-248
ISSN 2964- 5581 (media online)
https://utilityprojectsolution.org/ejournal/index.php/JuKSIT Penerbit : Cv. Utility Project Solutio

pada periode dan aplikasi lain. Secara keseluruhan, hasil menunjukkan bahwa kualitas benchmark tidak hanya ditentukan oleh jumlah model, tetapi juga oleh sumber label, cakupan lexicon, prosedur split, dan mode pelatihan. Transparansi pada seluruh tahapan tersebut penting agar hasil dapat direplikasi dan ditafsirkan secara tepat.

4. KESIMPULAN

Pelabelan InSet pada 10.071 ulasan menghasilkan 5,342 negatif, 1,019 netral, dan 3,710 positif, dengan coverage 93.80%. 4,360 ulasan yang tidak setuju dengan label penilaian, yang mencapai 43,29%, menunjukkan bahwa penilaian dan polaritas teks tidak dapat ditukarkan. Pada lima embrio, mean macro-F1 tertinggi dihasilkan oleh Linear SVM 0,711 ± 0,000, diikuti oleh Logistic Regression 0,709 ± 0,000. Menurut McNemar-Holm (p=0,5666), perbedaan keduanya tidak signifikan pada seed 42. Oleh karena itu, klaim model terbaik harus mempertimbangkan kedekatan skor dan biaya komputasi. Benchmark dua skema label dengan deduplikasi, split identik, lima seed, evaluasi berpasangan, dan audit performa-efisiensi adalah novelty yang dapat dipertahankan. Keterbatasan utamanya adalah InSet berasal dari domain microblog dan bukan anotasi manusia. Perbaikan domain lexicon, validasi manusia independen, penyempurnaan komprehensif transformer, dan evaluasi lintas aplikasi/periode semuanya harus diselesaikan sebelum pengembangan berikutnya. Secara praktis, baseline linear layak dipertimbangkan untuk penyaringan awal karena kinerja dan efisiensinya, tetapi hasil tidak boleh menjadi keputusan layanan langsung. Ulasan InSet tidak memiliki cakupan lexicon, prediksi tidak yakin, dan rating ketidaksepakatan. Hasil penelitian akademik menunjukkan bahwa sumber label dapat mengubah peringkat model. Oleh karena itu, untuk menilai validitas konstruk, penelitian lebih lanjut harus memeriksa hasil rating label, lexicon, dan subset label emas.

REFERENCES

[1] J. Dąbrowski, E. Letier, A. Perini, and A. Susi, "Analysing app reviews for software engineering: a systematic literature review," *Empir. Softw. Eng.*, vol. 27, no. 2, 2022, doi: 10.1007/s10664-021-10065-7.

[2] X. Wang, T. Zhang, Y. Tan, W. Shang, and Y. Li, "How to effectively mine app reviews concerning software ecosystem? A survey of review characteristics," Journal of Systems and Software, vol. 213, p. 112040, 2024, doi: 10.1016/j.jss.2024.112040.

[3] P. Rasappan, M. Premkumar, G. Sinha, and K. Chandrasekaran, "Transforming sentiment analysis for e-commerce product reviews: Hybrid deep learning model with an innovative term weighting and feature selection," Inf. Process. Manag., vol. 61, no. 3, p. 103654, May 2024, doi: 10.1016/j.ipm.2024.103654.

[4] P. Ataei, S. Regula, D. Staegemann, and S. Malgaonkar, "Filtering useful app reviews using Naive Bayes: Which Naive Bayes?," AI, vol. 5, no. 4, pp. 2237–2259, 2024, doi: 10.3390/ai5040110.

[5] A. Almansour, R. Alotaibi, and H. Alharbi, "Text-rating review discrepancy (TRRD): an integrative review and implications for research," Future Business Journal, vol. 8, no. 1, 2022, doi: 10.1186/s43093-022-00114-y.

[6] A. Iqbal, R. Amin, J. Iqbal, R. Alroobaea, A. Binmahfoudh, and M. Hussain, "Sentiment Analysis of Consumer Reviews Using Deep Learning," Sustainability, vol. 14, no. 17, p. 10844, 2022, doi: 10.3390/su141710844.

[7] A. Koufakou, "Deep learning for opinion mining and topic classification of course reviews," Educ. Inf. Technol. (Dordr)., vol. 29, no. 3, pp. 2973–2997, 2024, doi: 10.1007/s10639-023-11736-2.

[8] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," Journal of King Saud University - Computer and Information Sciences, vol. 36, no. 4, p. 102048, 2024, doi: 10.1016/j.jksuci.2024.102048.

[9] J. R. Jim, M. A. R. Talukder, P. Malakar, M. M. Kabir, K. Nur, and M. Mridha, "Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review," Natural Language Processing Journal, vol. 6, p. 100059, 2024, doi: 10.1016/j.nlp.2024.100059.

[10] Y. C. Hua, P. Denny, J. Wicker, and K. Taskova, "A systematic review of aspect-based sentiment analysis: domains, methods, and trends," Artif. Intell. Rev., vol. 57, no. 11, 2024, doi: 10.1007/s10462-024-10906-z.

[11] G. I. Winata, A. F. Aji, S. Cahyawijaya, R. Mahendra, F. Koto, and A. Romadhony, "NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages," in in Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Andreas Vlachos and Isabelle Augenstein, Eds., Association for Computational Linguistics, May 2023, pp. 815–834. doi: 10.18653/v1/2023.eacl-main.57.

[12] S. Cahyawijaya, H. Lovenia, A. F. Aji, G. Winata, B. Wilie, and F. Koto, "NusaCrowd: Open Source Initiative for Indonesian NLP Resources," in in Findings of the Association for Computational Linguistics: ACL, Anna Rogers,

Submitted: **29/05/2026**; Accepted: **20/06/2026**; Published:**22/07/2026**

Jurnal Kolaborasi
Sains dan Ilmu Terapan
Volume 5 Nomor 1 Juli-Desember 2026 | Page 237-248
ISSN 2964- 5581 (media online)
https://utilityprojectsolution.org/ejournal/index.php/JuKSIT Penerbit : Cv. Utility Project Solutio

Jordan Boyd-Graber, and Naoaki Okazaki, Eds., Association for Computational Linguistics, Jul. 2023, pp. 13745–13818. doi: 10.18653/v1/2023.findings-acl.868.

[13] B. Wilie et al., "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 843–857. doi: 10.18653/v1/2020.aacl-main.85.

[14] Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in in Proceedings *of the 28th International Conference on Computational Linguistics*, Donia Scott, Nuria Bel, and Chengqing Zong, Eds., International Committee on Computational Linguistics, Dec. 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[15] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language* Processing *(EMNLP)*, Alessandro Moschitti, Bo Pang, and Walter Daelemans, Eds., Stroudsburg, PA, USA: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751. doi: 10.3115/v1/D14-1181.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings *of the 2019 Conference of the North*, Jill Burstein, Christy Doran, and Thamar Solorio, Eds., Stroudsburg, PA, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[17] A. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational* Linguistics, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, Eds., Stroudsburg, PA, USA: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451. doi: 10.18653/v1/2020.acl-main.747.

[18] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.11692

[19] I. H. Setiawan, M. Rahardi, A. Aminuddin, and F. F. Abdulloh, "Sentiment Analysis of Tokopedia Application Reviews on Google Play Store Using BERT," in in, 2024, pp. 242–247. doi: 10.1109/ICITSI65188.2024.10929357.

[20] M. Idris, A. Rifai, and K. D. Tania, "Sentiment Analysis of Tokopedia App Reviews using Machine Learning and Word Embeddings," sinkron, vol. 9, no. 1, pp. 210–219, 2025, doi: 10.33395/sinkron.v9i1.14278.

[21] A. R. Aisy and G. Karyono, "Analisis Sentimen Terhadap Ulasan Google Play Store Aplikasi Lazada, Shopee, dan Tokopedia Menggunakan Algoritma IndoBERT," Building of Informatics, vol. 7, no. 3, pp. 2127–2135, 2025, doi: 10.47065/bits.v7i3.8745.

[22] R. Tajuddin, Y. Irawan, and R. R. Setiawan, "Classification of Sentiment Tokopedia and Shopee App Reviews on Google Playstore Using Naive Bayes," bit-Tech, vol. 8, no. 2, pp. 2348–2357, 2025, doi: 10.32877/bt.v8i2.3246.

[23] R. Wijianto, D. Pratmanto, A. Widayanto, and Ubaidilah, "Komparasi K-Nearest Neighbors (KNN) dan Naive Bayes pada Klasifikasi Sentimen Ulasan Aplikasi Tokopedia di Google Play Store," Informatics and Computer Engineering Journal, vol. 5, no. 2, pp. 75–80, 2025, doi: 10.31294/icej.v5i2.8939.

[24] F. D. Saputra and F. Budiman, "Comparison of Random Forest and LSTM for Tokopedia Sentiment Analysis," Journal of Applied Informatics and Computing, vol. 10, no. 1, pp. 630–639, 2026, doi: 10.30871/jaic.v10i1.12042.

[25] N. Alowidi, R. Alnanih, and N. Alsaleh, "Hybrid Deep Learning Approach for Automating App Review Classification: Advancing Usability Metrics Classification with an Aspect-Based Sentiment Analysis Framework," Computers, vol. 82, no. 1, pp. 949–976, 2024, doi: 10.32604/cmc.2024.059351.

[26] H. Murfi, Syamsyuriani, T. Gowandi, G. Ardaneswari, and S. Nurrohmah, "BERT-based combination of convolutional and recurrent neural network for indonesian sentiment analysis," Appl. Soft Comput., vol. 151, p. 111112, 2023, doi: 10.1016/j.asoc.2023.111112.

[27] Dhendra and V. Gayuh Utomo, "Benchmarking IndoBERT and Transformer Models for Sentiment Classification on Indonesian E-Government Service Reviews," Jurnal Transformatika, vol. 23, no. 1, pp. 86–95, Jul. 2025, doi: 10.26623/transformatika.v23i1.12095.

[28] C. Shaw, P. LaCasse, and L. Champagne, "Exploring emotion classification of indonesian tweets using large scale transfer learning via IndoBERT," Soc. Netw. Anal. Min., vol. 15, no. 1, 2025, doi: 10.1007/s13278-025-01439-6.

[29] Y. Wang, J. Wang, H. Zhang, X. Ming, and Q. Wang, "Better together: Automated app review analysis with deep multi-task learning," Inf. Softw. Technol., vol. 177, p. 107597, 2024, doi: 10.1016/j.infsof.2024.107597.

Submitted: **29/05/2026**; Accepted: **20/06/2026**; Published:**22/07/2026**

Jurnal Kolaborasi
Sains dan Ilmu Terapan
Volume 5 Nomor 1 Juli-Desember 2026 | Page 237-248
ISSN 2964- 5581 (media online)
https://utilityprojectsolution.org/ejournal/index.php/JuKSIT Penerbit : Cv. Utility Project Solutio

[30] N. Cassee, A. Agaronian, E. Constantinou, N. Novielli, and A. Serebrenik, "Transformers and meta-tokenization in sentiment analysis for software engineering," Empir. Softw. Eng., vol. 29, no. 4, 2024, doi: 10.1007/s10664-024-10468-2.

[31] F. Koto and G. Y. Rahmaningtyas, "Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs," in 2017 International Conference on Asian Language Processing (IALP), Singapore: IEEE, Dec. 2017, pp. 391–394. doi: 10.1109/IALP.2017.8300625.

[32] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Inf. Process. Manag., vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.

[33] T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," Neural Comput., vol. 10, no. 7, pp. 1895–1923, Oct. 1998, doi: 10.1162/089976698300017197.

[34] Y. Bestgen, "Please, Don't Forget the Difference and the Confidence Interval when Seeking for the State-of-the-Art Status," in in Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Jun. 2022, pp. 5956–5962. [Online]. Available: https://aclanthology.org/2022.lrec-1.640/

[35] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in in Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 1321–1330. [Online]. Available: https://proceedings.mlr.press/v70/guo17a.html

[36] T. Pires, E. Schlinger, and D. Garrette, "How Multilingual is Multilingual BERT?," in in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Anna Korhonen, David Traum, and Lluís Màrquez, Eds., Association for Computational Linguistics, Jul. 2019, pp. 4996–5001. doi: 10.18653/v1/P19-1493.

Submitted: **29/05/2026**; Accepted: **20/06/2026**; Published:**22/07/2026**