

**EFEKTIVITAS MODEL PEMROSESAN BAHASA ALAMI UNTUK ANALISIS
SENTIMEN TEKS BAHASA INDONESIA PADA DOMAIN LAYANAN PUBLIK:
SYSTEMATIC LITERATURE REVIEW 2021–2026**

R. Fais Putra Daviansyah¹, Galuh Boy Hertantyo², Priati Assiroj³

^{1,2,3}Politeknik Imigrasi dan Pemasyarakatan Indonesia

¹rfaisputradaviansyah@gmail.com

ABSTRACT

Public sentiment toward government digital services in Indonesia is increasingly expressed through application reviews and social media, yet the choice of natural language processing (NLP) model for Indonesian-language text remains inconsistent across studies. This study aims to identify the most effective NLP models for Indonesian sentiment analysis published in 2021–2026, with a particular focus on the public service domain. A systematic literature review was conducted following the PRISMA 2020 statement and the Kitchenham and Charters guidelines, covering Scopus, IEEE Xplore, ScienceDirect, SpringerLink, ACL Anthology, arXiv, SINTA, Garuda, and Google Scholar. Eighteen primary studies met the inclusion criteria and were synthesised descriptively and thematically along five research questions. The results show a shift from classical machine learning (Naive Bayes, SVM) in 2021–2023 toward Indonesian-specific transformer models (IndoBERT, IndoBERTweet) and hybrid architectures in 2024–2025. In within-study comparisons, IndoBERT and its variants generally outperformed classical and recurrent models, reaching accuracy or F1-scores above 0.90 on public service datasets such as Mobile JKN and Coretax. Large language models are emerging but have not consistently surpassed fine-tuned encoders on standard sentiment benchmarks. The main gaps are reliance on automatic lexicon- or rating-based labelling, accuracy reporting on imbalanced data without macro-F1, and the scarcity of aspect-based and shared public service benchmarks. IndoBERT-family models are recommended as the default baseline, together with manually annotated public service corpora and macro-F1 reporting.

Keywords: *sentiment analysis, IndoBERT, public service, natural language processing, systematic literature review*

ABSTRAK

Sentimen masyarakat terhadap layanan publik digital di Indonesia semakin banyak disampaikan melalui ulasan aplikasi dan media sosial, tetapi pemilihan model pemrosesan bahasa alami (NLP) untuk teks berbahasa Indonesia masih beragam antarstudi. Penelitian ini bertujuan mengidentifikasi model NLP yang paling efektif untuk analisis sentimen teks bahasa Indonesia pada publikasi 2021–2026 dengan fokus pada domain layanan publik. Tinjauan literatur sistematis dilakukan dengan mengacu pada PRISMA 2020 serta pedoman Kitchenham dan Charters pada basis data Scopus, IEEE Xplore, ScienceDirect, SpringerLink, ACL Anthology, arXiv, SINTA, Garuda, dan Google Scholar. Sebanyak delapan belas studi primer memenuhi kriteria inklusi dan disintesis secara deskriptif dan tematik berdasarkan lima pertanyaan penelitian. Hasil menunjukkan pergeseran dari *machine learning*

klasik (Naive Bayes, SVM) pada 2021–2023 ke model *transformer* khusus bahasa Indonesia (IndoBERT, IndoBERTweet) dan arsitektur hibrida pada 2024–2025. Dalam perbandingan di dalam studi yang sama, IndoBERT dan variannya umumnya mengungguli model klasik dan rekuren dengan akurasi atau F1 di atas 0,90 pada data layanan publik seperti Mobile JKN dan Coretax. *Large language model* mulai berkembang, tetapi belum konsisten melampaui *encoder* yang di-*fine-tune* pada tolok ukur sentimen standar. Celah utama meliputi ketergantungan pada pelabelan otomatis berbasis leksikon atau rating, pelaporan akurasi pada data tidak seimbang tanpa macro-F1, serta minimnya analisis berbasis aspek dan tolok ukur bersama untuk layanan publik. Model keluarga IndoBERT direkomendasikan sebagai *baseline* utama, disertai korpus layanan publik berlabel manual dan pelaporan macro-F1.

Kata Kunci: analisis sentimen, IndoBERT, layanan publik, pemrosesan bahasa alami, tinjauan literatur sistematis

A. Pendahuluan

Transformasi digital pemerintahan di Indonesia telah memindahkan banyak layanan publik ke kanal aplikasi seluler, mulai dari layanan kesehatan (Mobile JKN), energi (PLN Mobile dan MyPertamina), administrasi kependudukan (Identitas Kependudukan Digital), hingga perpajakan (Coretax). Seiring dengan itu, warga menyampaikan pengalaman, keluhan, dan apresiasi mereka melalui ulasan di Google Play maupun media sosial seperti X (Twitter). Analisis sentimen, yaitu cabang pemrosesan bahasa alami (*natural language processing/NLP*) yang mengklasifikasikan polaritas opini, menjadi pendekatan yang banyak dipakai untuk mengolah

umpan balik tersebut secara otomatis (Wankhade dkk., 2022).

Teks bahasa Indonesia di media sosial dan ulasan aplikasi memiliki karakteristik yang menyulitkan model: penggunaan bahasa tidak baku, singkatan, *slang*, sarkasme, serta campur kode dengan bahasa Inggris dan bahasa daerah. Pada periode awal, penelitian analisis sentimen berbahasa Indonesia banyak mengandalkan pendekatan berbasis leksikon seperti InSet (Koto & Rahmaningtyas, 2017) dan algoritma *machine learning* klasik seperti Naive Bayes dan *Support Vector Machine* (SVM). Kajian sistematis tentang analisis sentimen berbasis leksikon menunjukkan bahwa akurasi pendekatan ini umumnya tertahan di kisaran 90% (Fauziah dkk., 2021). Kemunculan arsitektur *transformer*

dan BERT (Devlin dkk., 2019) kemudian mendorong lahirnya model bahasa prapelatihan khusus bahasa Indonesia, yaitu IndoBERT dari IndoLEM (Koto dkk., 2020) dan IndoNLU (Wilie dkk., 2020), serta IndoBERTweet yang dilatih pada korpus Twitter Indonesia (Koto dkk., 2021). Tolok ukur NusaX (Winata dkk., 2023) turut memudahkan perbandingan model.

Perkembangan tersebut menimbulkan pertanyaan praktis bagi peneliti dan penyelenggara layanan publik: model mana yang sebenarnya paling efektif untuk teks bahasa Indonesia, dan apakah keunggulannya berlaku pula pada data layanan publik? Setiawan (2024) meninjau aplikasi analisis sentimen di Indonesia pada 2023–2024 dan menemukan dominasi Naive Bayes, SVM, dan KNN, tetapi tidak berfokus pada layanan publik dan tidak mencakup *large language model* (LLM). Fauziah dkk. (2021) hanya membahas pendekatan berbasis leksikon. Di sisi lain, banyak studi primer melaporkan akurasi sangat tinggi pada data yang tidak seimbang dan dilabeli secara otomatis, sehingga perbandingan lintas studi berisiko

menyesatkan bila tidak ditelaah secara kritis.

Berdasarkan celah tersebut, penelitian ini melakukan tinjauan literatur sistematis (*systematic literature review*/SLR) terhadap studi analisis sentimen teks bahasa Indonesia yang terbit pada 2021–2026 dengan prioritas pada domain layanan publik. Lima pertanyaan penelitian (*research question*/RQ) dirumuskan, yaitu: (RQ1) model apa yang paling banyak digunakan dan bagaimana distribusinya per tahun dan per kelompok model; (RQ2) model mana yang paling efektif berdasarkan akurasi dan F1-score; (RQ3) teknik pendukung apa yang digunakan, meliputi *preprocessing*, representasi fitur, dan penanganan data tidak seimbang; (RQ4) domain dan dataset apa yang diteliti, khususnya pada layanan publik; serta (RQ5) tantangan, celah penelitian, dan arah penelitian ke depan. Kontribusi penelitian ini adalah pemetaan model terkini, telaah kritis praktik evaluasi, dan rekomendasi pemilihan model untuk pemantauan opini layanan publik.

B. Metode Penelitian

Penelitian ini menggunakan desain tinjauan literatur sistematis yang mengacu pada pedoman Kitchenham dan Charters (2007) serta Kitchenham dkk. (2009), dengan pelaporan mengikuti pernyataan PRISMA 2020 (Page dkk., 2021). Protokol penelitian terdiri atas lima tahap: perumusan RQ, pencarian literatur, seleksi studi, penilaian kualitas, serta ekstraksi dan sintesis data.

1. Sumber Data dan Strategi

Pencarian

Pencarian dilakukan pada basis data internasional Scopus, IEEE Xplore, ScienceDirect, SpringerLink, ACM Digital Library, ACL Anthology, dan arXiv, serta sumber nasional SINTA, Garuda, dan Google Scholar untuk menjangkau jurnal nasional terakreditasi. *Search string* utama untuk basis data berbahasa Inggris adalah ("sentiment analysis" OR "opinion mining") AND ("Indonesian" OR "Bahasa Indonesia") AND ("machine learning" OR "deep learning" OR "BERT" OR "IndoBERT" OR "transformer" OR "lexicon"), dengan batas tahun terbit 2021–2026. Untuk sumber nasional digunakan padanan berbahasa Indonesia

dengan frasa "analisis sentimen". Pencarian spesifik layanan publik ditambahkan dengan kata kunci nama aplikasi, antara lain "Mobile JKN", "BPJS", "PLN Mobile", "MyPertamina", "PeduliLindungi", "IKD", dan "Coretax". Penelusuran dilengkapi dengan *backward snowballing* pada daftar pustaka studi terpilih (Wohlin, 2014).

2. Kriteria Inklusi dan Eksklusi

Studi diikutsertakan apabila (1) merupakan studi primer analisis sentimen atau emosi; (2) menggunakan data teks bahasa Indonesia, termasuk teks informal dan campur kode; (3) terbit pada 2021–2026; (4) melaporkan model serta sedikitnya satu metrik kuantitatif (akurasi, presisi, *recall*, atau F1-score); (5) terbit pada jurnal, prosiding, atau *preprint* yang dapat diakses; dan (6) ditulis dalam bahasa Indonesia atau Inggris. Studi dikeluarkan apabila tidak menggunakan bahasa Indonesia, bukan tugas analisis sentimen (misalnya hanya pemodelan topik), tidak memiliki metrik evaluasi, merupakan duplikasi, merupakan literatur sekunder, atau teks lengkapnya tidak tersedia.

3. Seleksi Studi dan Penilaian

Kualitas

Seleksi mengikuti tahapan PRISMA 2020 sebagaimana ditunjukkan pada Gambar 1. Setelah duplikat dihapus, pencarian basis data dan *backward snowballing* menghasilkan 46 studi kandidat. Pada tahap penyaringan, enam artikel

tinjauan dikeluarkan dan hanya digunakan sebagai kajian terkait. Sebanyak 40 studi kemudian dinilai kelayakannya berdasarkan teks lengkap dan metadata bibliografis. Sebanyak 22 studi dikeluarkan dengan alasan yang dirinci pada Gambar 1. Dengan demikian, 18 studi primer disertakan dalam sintesis.

Tahap	Proses	Studi dikeluarkan
Identifikasi	Studi kandidat dari Scopus, IEEE Xplore, ScienceDirect, SpringerLink, ACM DL, ACL Anthology, arXiv, SINTA, Garuda, Google Scholar, dan backward snowballing setelah duplikat dihapus (n = 46)	–
Penyaringan	Judul dan abstrak disaring (n = 46)	Literatur sekunder/tinjauan, digunakan sebagai kajian terkait (n = 6)
Kelayakan	Teks lengkap dan metadata bibliografis dinilai (n = 40)	Dikeluarkan (n = 22): metadata tidak dapat diverifikasi (16); tanpa metrik kuantitatif (4); teks lengkap tidak diperoleh (1); evaluasi multibahasa (1)
Disertakan	Studi primer yang disintesis (n = 18)	–

Gambar 1 Diagram Alir Seleksi Studi Berdasarkan PRISMA 2020

Kualitas setiap studi dinilai dengan delapan pertanyaan (QA1–QA8) yang mencakup kejelasan tujuan, deskripsi sumber data, kesesuaian *preprocessing* untuk bahasa Indonesia, metode pelabelan, pengaturan eksperimen, kecukupan metrik evaluasi (tidak hanya akurasi pada data tidak seimbang), keberadaan pembanding atau validasi silang, serta pembahasan keterbatasan. Setiap butir diberi skor 1 (ya), 0,5 (sebagian), atau 0 (tidak), sehingga skor maksimum adalah 8.

Studi dengan skor di bawah 4 dikeluarkan; skor 4–6 dikategorikan sedang dan di atas 6 dikategorikan tinggi. Hasil penilaian menunjukkan 12 studi berkualitas tinggi dan 6 studi berkualitas sedang, sehingga tidak ada studi yang dikeluarkan pada tahap ini.

4. Ekstraksi dan Sintesis Data

Data yang diekstraksi dari setiap studi meliputi penulis dan tahun, domain atau objek, sumber dan ukuran data, metode pelabelan, tahap *preprocessing*, representasi fitur,

model, penanganan data tidak seimbang, serta metrik dan hasil. Sintesis dilakukan secara kuantitatif-deskriptif (frekuensi model per tahun dan per domain) dan naratif-tematik untuk menjawab RQ3–RQ5. Karena dataset, skema pelabelan, dan jumlah kelas berbeda antarstudi, perbandingan efektivitas diutamakan pada perbandingan di dalam studi yang sama (*within-study comparison*), bukan perbandingan angka mentah antarstudi.

C. Hasil Penelitian dan Pembahasan

1. Karakteristik Studi Terpilih

Proses seleksi menghasilkan 18 studi primer yang memenuhi seluruh kriteria. Ringkasan hasil ekstraksi disajikan pada Tabel 1. Sebanyak 10 studi (55,6%) secara langsung meneliti aplikasi layanan publik, sedangkan delapan studi lainnya meneliti domain umum, pendidikan, opini publik, atau tolok ukur yang relevan untuk pemilihan model. Distribusi terbitan meningkat tajam pada 2025 (8 studi atau 44,4%), sementara tidak ada studi terbitan 2026 yang memenuhi kriteria hingga pencarian dilakukan.

**Tabel 1 Ringkasan Ekstraksi Studi Primer Analisis Sentimen
Teks Bahasa Indonesia (2021–2026)**

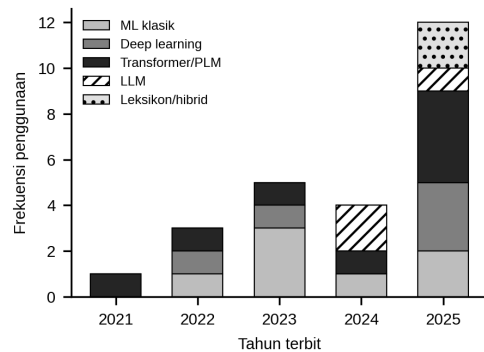
No	Penulis (Tahun)	Domain/Objek	Data	Model	Hasil utama
1	Koto dkk. (2021)	Twitter Indonesia (umum)	Korpus tweet Indonesia; tugas sentimen Twitter	IndoBERTweet vs IndoBERT, mBERT	IndoBERTweet unggul atas IndoBERT & mBERT pada tugas Twitter
2	Yunanto & Yulianto (2022)	PeduliLindungi (LP)	Twitter	NB vs SVM	SVM 91%; NB 90%
3	Mandhasiya dkk. (2022)	Opini publik Pilpres 2024	Komentar YouTube	BERT + LSTM/CNN/CNN-LSTM	BERT-LSTM akurasi 87,83% (unggul atas BERT murni)
4	Ellyanti dkk. (2023)	PeduliLindungi (LP)	Twitter, Jun–Des 2021	NB (10-fold CV)	Akurasi 95,86%; presisi 96,99%; recall 94,12%
5	Maulana dkk. (2023)	MyPertamina (LP)	Google Play, 5.722 ulasan	NBC	Akurasi ±94%
6	Imaduddin dkk. (2023)	Aplikasi layanan kesehatan	Google Play, 9.310 ulasan (3 aplikasi)	IndoBERT vs Fast Large-Margin, BiLSTM	IndoBERT akurasi 96%, F1 95%; FLM 92,33%; BiLSTM 80,88%
7	Komarudin & Hilda (2024)	Identitas Kependudukan Digital (LP)	Google Play, 1.000 ulasan	NB	Akurasi 82,23%; recall 94,02%
8	Yulianti & Nissa (2024)	Ulasan pelanggan (ABSA)	Ulasan pelanggan berbahasa Indonesia	IndoBERT single-sentence & sentence-pair	F1 +2,2% vs mBERT-aux; +23,6% vs Word2Vec-CNN-XGBoost

No	Penulis (Tahun)	Domain/Objek	Data	Model	Hasil utama
9	Cahyawijaya dkk. (2024)	Benchmark NLU Indonesia	Tugas NLU termasuk sentimen	Cendol (LLM instruction-tuned)	Varian terbesar unggul $\pm 20\%$ weighted F1 atas LLM lain
10	Owen dkk. (2024)	Benchmark NusaX-Senti	NusaX	Komodo-7B-Instruct	NusaX-Senti 79,3 vs GPT-3.5 47,2 (dilaporkan pengembang)
11	Tarwoto dkk. (2025)	Mobile JKN (BPJS Kesehatan) (LP)	Google Play, 3 kelas	IndoBERT	Akurasi rata-rata 97,28%; terbaik 98,27%
12	Tamami dkk. (2025)	Mobile JKN (LP)	Ulasan aplikasi (tidak seimbang)	LSTM + SMOTE	Akurasi 88%; F1 89% (dengan SMOTE)
13	Asri dkk. (2025)	PLN Mobile (LP)	Google Play, 1.000 ulasan; label InSet	IndoBERT + leksikon InSet	Validasi $\pm 85-86\%$
14	Ayomi dkk. (2025)	PLN Mobile (LP)	Ulasan pengguna	NB vs LSTM	NB 70,89% (F1 0,6964); LSTM 98,02% (F1 0,9800)
15	Rizkia dkk. (2025)	Coretax (DJP) (LP)	8.035 tweet; 5 skema pelabelan	IndoBERT	F1 tertinggi 0,9802 (label IndoBERTtweet; bias kelas netral)
16	Dermawan & Ayunda (2025)	Coretax (DJP) (LP)	Media sosial X	NB, SVM, LSTM + SMOTE	SVM (RBF) terbaik: akurasi 90,70%; F1 90,66%
17	Jazuli dkk. (2025)	Umpan balik mahasiswa (ABSA)	10.000 ulasan mahasiswa	IndoBERT ABSA	Aspek: akurasi 0,973 (F1 0,952); sentimen: akurasi 0,979 (F1 0,974)
18	Nasution dkk. (2025)	Tweet Indonesia (sentimen & emosi)	Tweet; 22 LLM open-source + GPT-4	LLM zero-shot	GPT-4 macro-F1 0,84 (emosi) vs IndoBERT-Large 0,791

Keterangan: LP = aplikasi/layanan publik; NB = Naive Bayes; SVM = Support Vector Machine; ABSA = aspect-based sentiment analysis.

2. Model yang Dominan (RQ1)

Gambar 2 memperlihatkan frekuensi penggunaan kelompok model per tahun; satu studi dapat menggunakan lebih dari satu kelompok model. Secara keseluruhan, model *transformer/pretrained language model* (PLM) digunakan pada 8 studi (44,4%), *machine learning* klasik pada 7 studi (38,9%), *deep learning* non-*transformer* pada 5 studi (27,8%), LLM pada 3 studi (16,7%), dan leksikon sebagai komponen hibrida atau pelabelan pada 2 studi (11,1%).



Gambar 2 Frekuensi Penggunaan Kelompok Model per Tahun (n = 18 studi)

Pola temporal memperlihatkan tiga fase. Pada 2021–2023, studi layanan publik didominasi Naive Bayes dan SVM dengan fitur TF-IDF, misalnya pada PeduliLindungi (Yunanto & Yulianto, 2022; Ellyanti

dkk., 2023) dan MyPertamina (Maulana dkk., 2023). Pada 2023–2024, IndoBERT mulai digunakan secara luas, termasuk pada aplikasi layanan kesehatan (Imaduddin dkk., 2023) dan analisis sentimen berbasis aspek (Yulianti & Nissa, 2024). Pada 2024–2025, penelitian bergeser ke IndoBERT yang di-*fine-tune* pada data layanan publik (Tarwoto dkk., 2025; Rizkia dkk., 2025), kombinasi IndoBERT dengan leksikon (Asri dkk., 2025), penanganan data tidak seimbang dengan SMOTE (Tamami dkk., 2025), serta eksplorasi LLM (Cahyawijaya dkk., 2024; Owen dkk., 2024; Nasution dkk., 2025). Temuan ini sejalan dengan Setiawan (2024) yang memperkirakan pergeseran dari algoritma klasik ke *deep learning*.

3. Efektivitas Model (RQ2)

Perbandingan di dalam studi yang sama memberikan bukti paling kuat tentang efektivitas relatif. Imaduddin dkk. (2023) melaporkan IndoBERT mencapai akurasi 96% dan F1 95% pada ulasan aplikasi kesehatan, melampaui Fast Large-Margin (92,33%) dan BiLSTM (80,88%). Yulianti dan Nissa (2024) menunjukkan IndoBERT dengan pendekatan *sentence-pair* unggul 2,2

poin F1 atas mBERT dan 23,6 poin atas Word2Vec-CNN-XGBoost. Koto dkk. (2021) menunjukkan bahwa IndoBERTweet mengungguli IndoBERT dan mBERT pada tugas-tugas Twitter Indonesia, yang menegaskan manfaat prapelatihan pada domain media sosial. Mandhasiya dkk. (2022) menemukan bahwa representasi BERT yang dipadukan dengan LSTM (akurasi 87,83%) lebih baik daripada BERT murni pada komentar YouTube terkait pemilihan presiden.

Pada data layanan publik, IndoBERT mencapai akurasi rata-rata 97,28% dengan nilai terbaik 98,27% pada ulasan Mobile JKN (Tarwoto dkk., 2025) dan F1 hingga 0,9802 pada tweet Coretax (Rizkia dkk., 2025). Namun, keunggulan *transformer* tidak bersifat mutlak. Ayomi dkk. (2025) melaporkan LSTM mencapai akurasi 98,02% (F1 0,9800) pada ulasan PLN Mobile, jauh di atas Naive Bayes (70,89%). Hal ini menunjukkan bahwa kinerja sangat dipengaruhi oleh karakteristik data, skema pelabelan, dan keseimbangan kelas. Tabel 2 merangkum rentang kinerja setiap kelompok model.

Tabel 2 Perbandingan Kelompok Model pada Studi Terpilih

Kelompok model	Contoh	Kinerja yang dilaporkan	Kekuatan	Keterbatasan
ML klasik	NB, SVM, Fast Large-Margin	Akurasi 70,89–95,86%	Ringan, cepat, mudah diinterpretasi	Lemah pada konteks dan campur kode; akurasi dapat menyesatkan pada data tidak seimbang
Deep learning	LSTM, BiLSTM, BERT-LSTM	Akurasi 80,88–98,02%	Menangkap urutan kata	Sensitif terhadap embedding dan ukuran data
Transformer/PL M	IndoBERT, IndoBERTweet	Akurasi $\pm$ 85–98,27%; F1 0,95–0,98	Pemahaman konteks terbaik untuk bahasa Indonesia	Memerlukan GPU dan fine-tuning
LLM	GPT-4, Cendol, Komodo	NusaX 47,2–79,3; macro-F1 emosi 0,84	Zero-/few-shot tanpa data latih besar	Belum konsisten melampaui encoder; biaya dan privasi; sebagian angka dilaporkan pengembang

Untuk LLM, bukti yang tersedia masih beragam. Pada tolok ukur NusaX, Komodo-7B-Instruct melaporkan skor 79,3, jauh di atas GPT-3.5 (47,2), tetapi angka ini dilaporkan oleh pengembang model (Owen dkk., 2024). Cendol, LLM yang di-*instruction-tune* untuk bahasa Indonesia dan bahasa daerah, melampaui LLM lain sekitar 20% *weighted* F1 pada tugas pemahaman bahasa (Cahyawijaya dkk., 2024). Evaluasi independen BHASA menunjukkan bahwa model *encoder* yang di-*fine-tune* masih unggul atas GPT-4 *zero-shot* pada sentimen NusaX (Leong dkk., 2023), sedangkan pada klasifikasi emosi GPT-4 mencapai macro-F1 0,84 dan melampaui IndoBERT-Large (0,791) (Nasution dkk., 2025). Dengan demikian, LLM menjanjikan untuk

skenario *zero-shot*, tetapi belum konsisten menggantikan IndoBERT.

4. Teknik Pendukung (RQ3)

Tahap *preprocessing* yang umum ditemukan meliputi pembersihan teks, *case folding*, tokenisasi, normalisasi kata tidak baku menggunakan kamus *slang*, penghapusan *stopword*, dan *stemming* dengan Sastrawi. Representasi fitur bergeser dari TF-IDF (ML klasik) dan Word2Vec/FastText (*deep learning*) ke *embedding* kontekstual *transformer*. Pelabelan otomatis berbasis leksikon InSet atau rating bintang masih sering digunakan (Asri dkk., 2025). Rizkia dkk. (2025) memperlihatkan bahwa skema pelabelan berpengaruh besar: F1 tertinggi diperoleh dengan label yang dihasilkan IndoBERTweet, tetapi label

tersebut bias terhadap kelas netral. Untuk data tidak seimbang, SMOTE merupakan teknik yang paling sering digunakan, misalnya pada ulasan Mobile JKN dengan LSTM (F1 89%; Tamami dkk., 2025) dan tweet Coretax dengan SVM (F1 90,66%; Dermawan & Ayunda, 2025).

5. Domain dan Dataset Layanan Publik (RQ4)

Sepuluh studi layanan publik mencakup enam objek, yaitu PeduliLindungi (kesehatan masyarakat), Mobile JKN (jaminan kesehatan), PLN Mobile (kelistrikan), MyPertamina (energi), Identitas Kependudukan Digital (administrasi kependudukan), dan Coretax (perpajakan). Sumber data didominasi ulasan Google Play dan tweet, dengan ukuran data yang relatif kecil, yakni ribuan hingga belasan ribu teks per studi, misalnya 5.722 ulasan MyPertamina (Maulana dkk., 2023) dan 8.035 tweet Coretax (Rizkia dkk., 2025). Analisis sentimen berbasis aspek (ABSA) yang mampu memetakan keluhan per aspek layanan, seperti pendaftaran, verifikasi, pembayaran, dan antarmuka, telah terbukti efektif pada ulasan pelanggan dan umpan balik pendidikan (Yulianti & Nissa, 2024;

Jazuli dkk., 2025), tetapi belum banyak diterapkan pada aplikasi layanan publik.

6. Tantangan dan Arah Penelitian (RQ5)

Sintesis mengidentifikasi lima tantangan utama. Pertama, ketergantungan pada pelabelan otomatis berbasis leksikon atau rating menimbulkan bias label, sehingga diperlukan anotasi manual dengan pengukuran kesepakatan antaranotator. Kedua, banyak studi hanya melaporkan akurasi pada data yang tidak seimbang; akurasi di atas 95% pada kondisi tersebut perlu dibaca dengan hati-hati, dan macro-F1 lebih tepat digunakan. Ketiga, bahasa informal, sarkasme, dan campur kode masih menjadi sumber kesalahan utama. Keempat, ABSA untuk layanan publik masih jarang. Kelima, belum tersedia korpus layanan publik berlabel yang dibagikan secara terbuka sebagai tolok ukur bersama, sehingga hasil antarstudi sulit dibandingkan. Penerapan LLM juga perlu memperhatikan biaya, reproduktibilitas, dan perlindungan data warga.

Tinjauan ini memiliki keterbatasan: sebagian angka kinerja

diambil dari abstrak dengan metrik yang tidak seragam, perbedaan dataset dan skema pelabelan membatasi perbandingan antarstudi, dan sebagian angka LLM dilaporkan sendiri oleh pengembang model.

D. Kesimpulan

Berdasarkan tinjauan sistematis terhadap 18 studi primer yang terbit pada 2021–2026, model *transformer* khusus bahasa Indonesia, terutama IndoBERT dan varian IndoBERTweet, merupakan pendekatan yang paling efektif dan paling relevan untuk analisis sentimen teks bahasa Indonesia, termasuk pada domain layanan publik. Model ini secara konsisten mengungguli *machine learning* klasik dan umumnya mengungguli *deep learning non-transformer* dalam perbandingan di dalam studi yang sama, dengan akurasi atau F1 di atas 0,90 pada data Mobile JKN dan Coretax. Penelitian bergeser dari Naive Bayes dan SVM pada 2021–2023 menuju IndoBERT, arsitektur hibrida, dan penanganan data tidak seimbang pada 2024–2025. LLM mulai menjadi tren yang menjanjikan, tetapi belum konsisten melampaui *encoder* yang di-*fine-tune*.

Penelitian selanjutnya disarankan untuk menjadikan IndoBERT atau IndoBERTweet sebagai *baseline* utama, melaporkan macro-F1 di samping akurasi, membangun korpus keluhan layanan publik berlabel manual yang terbuka, mengembangkan ABSA untuk aspek-aspek layanan, serta mengevaluasi LLM berbahasa Indonesia secara independen pada data layanan publik.

DAFTAR PUSTAKA

- Asri, Y., Kuswardani, D., Suliyanti, W. N., Manullang, Y. O., & Ansyari, A. R. (2025). Sentiment analysis based on Indonesian language lexicon and IndoBERT on user reviews PLN mobile application. *Indonesian Journal of Electrical Engineering and Computer Science*, 38(1), 677–688. <https://doi.org/10.11591/ijeecs.v38.i1.pp677-688>
- Ayomi, J. M., Vitianingsih, A. V., Kristyawan, Y., Maukar, A. L., & Widiartin, T. (2025). Sentiment analysis of user reviews for the PLN Mobile application using Naïve Bayes and long short-term memory. *Journal of Information Systems and Informatics*, 7(4), 3849–3873. <https://doi.org/10.63158/journalisi.v7i4.1342>
- Cahyawijaya, S., Lovenia, H., Koto, F., Putri, R. A., Dave, E., Lee, J., ... Fung, P. (2024). Cendol: Open instruction-tuned generative large language models for Indonesian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational*

- Linguistics* (Volume 1: Long Papers) (pp. 14899–14914). <https://doi.org/10.18653/v1/2024.acl-long.796>
- Dermawan, S., & Ayunda, A. T. (2025). Sentiment analysis of Coretax on social media X using Naive Bayes, SVM, and LSTM for service improvement. *Journal of Applied Informatics and Computing*, 9(6), 3177–3190. <https://doi.org/10.30871/jaic.v9i6.11063>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). <https://doi.org/10.18653/v1/N19-1423>
- Ellyanti, L., Ruldeviyani, Y., Pradana, L. E., & Harjanto, A. (2023). Sentiment analysis of Twitter users to the PeduliLindungi using Naïve Bayes algorithm. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 7(2), 414–421. <https://jurnal.iaii.or.id/index.php/RESTI/article/view/4684>
- Fauziah, Y., Yuwono, B. D., & Aribowo, A. S. (2021). Lexicon based sentiment analysis in Indonesia languages: A systematic literature review. *RSF Conference Series: Engineering and Technology*, 1(1), 363–367. <https://doi.org/10.31098/cset.v1i1.397>
- Imaduddin, H., A'la, F. Y., & Nugroho, Y. S. (2023). Sentiment analysis in Indonesian healthcare applications using IndoBERT approach. *International Journal of Advanced Computer Science and Applications*, 14(8). <https://doi.org/10.14569/IJACSA.2023.0140813>
- Jazuli, A., Widowati, W., & Kusumaningrum, R. (2025). Optimizing aspect-based sentiment analysis using BERT for comprehensive analysis of Indonesian student feedback. *Applied Sciences*, 15(1), 172. <https://doi.org/10.3390/app15010172>
- Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology*, 51(1), 7–15. <https://doi.org/10.1016/j.infsof.2008.09.009>
- Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering* (Technical Report EBSE-2007-01). Keele: Keele University and Durham University.
- Komarudin, A., & Hilda, A. M. (2024). Analisis sentimen ulasan aplikasi Identitas Kependudukan Digital pada Play Store menggunakan metode Naïve Bayes. *Computer Science (CO-SCIENCE)*, 4(1), 28–36. <https://doi.org/10.31294/coscience.v4i1.2955>
- Koto, F., Lau, J. H., & Baldwin, T. (2021). IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 10660–10668). <https://doi.org/10.18653/v1/2021.emnlp-main.833>

- Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 757–770).
<https://doi.org/10.18653/v1/2020.coling-main.66>
- Koto, F., & Rahmaningtyas, G. Y. (2017). InSet lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs. In *2017 International Conference on Asian Language Processing (IALP)* (pp. 391–394).
<https://doi.org/10.1109/IALP.2017.8300625>
- Leong, W. Q., Ngui, J. G., Susanto, Y., Rengarajan, H., Sarveswaran, K., & Tjhi, W. C. (2023). *BHASA: A holistic Southeast Asian linguistic and cultural evaluation suite for large language models* (arXiv:2309.06085).
<https://arxiv.org/abs/2309.06085>
- Mandhasiya, D. G., Murfi, H., Bustamam, A., & Anki, P. (2022). Evaluation of machine learning performance based on BERT data representation with LSTM model to conduct sentiment analysis in Indonesian for predicting voices of social media users in the 2024 Indonesia presidential election. In *2022 5th International Conference on Information and Communications Technology (ICOIACT)* (pp. 441–446).
<https://doi.org/10.1109/ICOIACT55506.2022.9972206>
- Maulana, R., Voutama, A., & Ridwan, T. (2023). Analisis sentimen ulasan aplikasi MyPertamina pada Google Play Store menggunakan algoritma NBC. *Jurnal Teknologi Terpadu*, 9(1), 42–48.
<https://doi.org/10.54914/jtt.v9i1.609>
- Nasution, A. H., Onan, A., Murakami, Y., Monika, W., & Hanafiah, A. (2025). Benchmarking open-source large language models for sentiment and emotion classification in Indonesian tweets. *IEEE Access*, 13, 94009–94025.
<https://doi.org/10.1109/ACCESS.2025.3574629>
- Owen, L., Tripathi, V., Kumar, A., & Ahmed, B. (2024). *Komodo: A linguistic expedition into Indonesia's regional languages* (arXiv:2403.09362).
<https://arxiv.org/abs/2403.09362>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71.
<https://doi.org/10.1136/bmj.n71>
- Rizkia, A. S., Wufron, W., & Roji, F. F. (2025). Analisis sentimen Coretax: Perbandingan pelabelan data manual, transformers-based, dan lexicon-based pada performa IndoBERT. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 5(3), 1037–1048.
<https://doi.org/10.57152/malcom.v5i3.2151>
- Setiawan, B. (2024). A review of sentiment analysis applications in Indonesia between 2023–2024. *JIEET (Journal of Information Engineering and Educational Technology)*, 8(2), 71–83.
<https://doi.org/10.26740/jieet.v8n2.p71-83>
- Tamami, G., Triyanto, W. A., & Muzid, S. (2025). Sentiment analysis Mobile JKN reviews using SMOTE based LSTM. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 19(1), 13–24.

- <https://doi.org/10.22146/ijccs.101910>
- Tarwoto, Nugroho, R., Azka, N., & Graha, W. S. R. (2025). Analisis sentimen ulasan aplikasi Mobile JKN di Google PlayStore menggunakan IndoBERT. *Jurnal JTik (Jurnal Teknologi Informasi dan Komunikasi)*, 9(2), 495–505. <https://doi.org/10.35870/jtik.v9i2.3340>
- Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7), 5731–5780. <https://doi.org/10.1007/s10462-022-10144-1>
- Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., ... Purwarianti, A. (2020). 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* (pp. 843–857). Association for Computational Linguistics.
- Winata, G. I., Aji, A. F., Cahyawijaya, S., Mahendra, R., Koto, F., Romadhony, A., ... Ruder, S. (2023). NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 815–834). <https://doi.org/10.18653/v1/2023.eacl-main.57>
- Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering* (Article 38). <https://doi.org/10.1145/2601248.2601268>
- Yulianti, E., & Nissa, N. K. (2024). ABSA of Indonesian customer reviews using IndoBERT: Single-sentence and sentence-pair classification approaches. *Bulletin of Electrical Engineering and Informatics*, 13(5), 3579–3589. <https://doi.org/10.11591/eei.v13i5.8032>
- Yunanto, I., & Yulianto, S. (2022). Twitter sentiment analysis PeduliLindungi application using Naïve Bayes and support vector machine. *Jurnal Teknik Informatika (JUTIF)*, 3(4), 807–814. <https://doi.org/10.20884/1.jutif.2022.3.4.292>