Evaluasi Bi-LSTM+FastText, IndoBERT, SVM, dan Naïve Bayes dengan Uji Signifikansi Statistik

John Aldo Metanfanuan[1], Bambang Nurdewanto[2]
Sistem Informasi, Universitas Merdeka Malang
e-mail: johnaldometanfanuan821@gmail.com[1], nurdewa@unmer.ac.id[2]

Kata Kunci:

Analisis sentimen
IndoBERT
Bi-LSTM
FastText
McNemar's Test
Cohen's Kappa
AI generatif

ABSTRAK

Penelitian ini membandingkan kinerja empat model klasifikasi sentimen Naïve Bayes+TF-IDF, Support Vector Machine (SVM)+TF-IDF, Bidirectional Long Short-Term Memory dengan pre-trained embedding FastText (Bi-LSTM+FastText), dan IndoBERT pada domain ulasan aplikasi Claude AI berbahasa Indonesia. Sebanyak 1.566 ulasan dikumpulkan dari Google Play Store melalui teknik web scraping dan dilabeli secara manual oleh tiga annotator independen dengan jumlah data yang disetarakan, dengan validitas inter-annotator agreement diukur menggunakan Cohen's Kappa (rata-rata K=0,9101; hampir sempurna). Seluruh model dievaluasi pada partisi data uji yang identik (n=314) menggunakan metrik akurasi, presisi, recall, dan weighted F1-score, serta diuji signifikansi statistiknya melalui McNemar's Test dan effect size Cohen's h. Hasil menunjukkan IndoBERT meraih Weighted F1-Score tertinggi (0,5159) meski akurasinya bukan yang tertinggi (0,5223), dan menjadi satu-satunya model yang mampu mengklasifikasikan kelas netral secara bermakna (F1=0,2143). Uji McNemar pada keenam pasangan model tidak menemukan perbedaan yang signifikan secara statistik (seluruh p>0,05; effect size kecil), sehingga hipotesis nol tidak dapat ditolak.

Keyword:

Sentiment analysis
IndoBERT
Bi-LSTM
FastText
McNemar's Test
Cohen's Kappa
Generative AI

ABSTRACT

This study compares the performance of four sentiment classification models Naïve Bayes+TF-IDF, Support Vector Machine (SVM)+TF-IDF, Bidirectional Long Short-Term Memory with pre-trained FastText embeddings (Bi-LSTM+FastText), and IndoBERT on the domain of Indonesian-language Claude AI application reviews. A dataset of 1,566 reviews was collected from Google Play Store via web scraping and manually annotated by three independent annotators with an equalized data count, with inter-annotator agreement validated using Cohen's Kappa (mean K=0.9101; almost perfect agreement). All models were evaluated on an identical test partition (n=314) using accuracy, precision, recall, and weighted F1-score, with statistical significance verified through McNemar's Test and Cohen's h effect size. Results show that IndoBERT achieved the highest Weighted F1-Score (0.5159) despite not having the highest accuracy (0.5223), and was the only model capable of meaningfully classifying the neutral class (F1=0.2143). McNemar's test across all six model pairs found no statistically significant differences (all p>0.05; small effect sizes), and the null hypothesis could not be rejected.

Copyright © SENASIF 2026
ISSN: 2598-0076

I. PENDAHULUAN

Perkembangan ekosistem kecerdasan buatan generatif yang pesat telah melahirkan berbagai asisten percakapan berbasis large language model (LLM), salah satunya Claude AI yang dikembangkan Anthropic dengan pendekatan Constitutional AI (CAI) sebagai fondasi pelatihan modelnya [1]. Di Indonesia, tren adopsi teknologi AI generatif mencatat akselerasi signifikan, tercermin dari posisi Indonesia sebagai penyumbang traffic terbesar ketiga secara global ke berbagai platform AI, dengan kontribusi 5,60% dari total kunjungan dunia [2]. Fenomena ini mendorong melimpahnya ulasan berbahasa Indonesia di platform Google Play Store yang menyimpan opini dan penilaian autentik pengguna [3], membentuk sumber data bernilai tinggi untuk memetakan persepsi publik melalui analisis sentimen.

Namun, melimpahnya volume ulasan menciptakan tantangan tersendiri. Analisis manual tidak lagi efisien dan rentan bias subjektif [4], sementara pendekatan machine learning konvensional berbasis TF-IDF memiliki keterbatasan struktural dalam menangkap konteks semantik [5]. Tantangan berlapis dijumpai pada teks berbahasa Indonesia yang kaya variasi linguistik slang, code-mixing, dan singkatan khas platform digital [6]. Lebih jauh, ulasan terhadap aplikasi AI generatif mengandung dimensi penilaian yang jauh lebih kompleks dibanding ulasan aplikasi konvensional, mencakup kualitas penalaran, konsistensi, dan aspek etika model.

Tinjauan literatur mengidentifikasi kesenjangan penelitian yang signifikan: belum ada kajian yang secara spesifik menelaah sentimen pengguna Indonesia terhadap Claude AI. Penelitian oleh Illahi et al. [7] membuktikan efektivitas kombinasi Bi-LSTM dan IndoBERT pada skenario data terbatas, mencapai F1-score 71% pada data validasi. Nngrum et al. [8] menemukan bahwa arsitektur Transformer (BERT) lebih unggul dibandingkan RNN-LSTM pada analisis sentimen aplikasi perbankan setelah optimasi tokenisasi. Namun demikian, penelitian terdekat pada domain LLM oleh Mahendra et al. [9] pada aplikasi DeepSeek AI hanya mencapai macro-average F1-score 0,33, mengungkap tantangan besar dalam menangani imbalanced data pada domain ini. Sebagian besar penelitian terdahulu juga membandingkan model hanya berdasarkan selisih nilai metrik tanpa pengujian signifikansi statistik formal, sebuah praktik yang berisiko menghasilkan kesimpulan prematur, khususnya pada kondisi dataset berskala kecil dan tidak seimbang yang umum dijumpai pada domain riset yang masih baru.

Penelitian ini bertujuan mengisi kesenjangan tersebut dengan tiga kontribusi utama. Pertama, membangun dataset ulasan Claude AI berbahasa Indonesia dengan pelabelan manual terstandarisasi dan validasi Cohen's Kappa. Kedua, membandingkan empat model klasifikasi secara setara Bi-LSTM dengan pre-trained embedding FastText, IndoBERT, SVM+TF-IDF, dan Naïve Bayes+TF-IDF menggunakan partisi data yang identik, dengan penyetaraan kondisi awal Bi-LSTM melalui embedding FastText Bahasa Indonesia [10] agar perbandingan dengan IndoBERT bersifat adil secara metodologis. Ketiga, melaporkan uji signifikansi statistik formal melalui McNemar's Test [11] dan

Copyright © SENASIF 2026
ISSN: 2598-0076

effect size Cohen's h [12], melengkapi metrik evaluasi konvensional yang selama ini menjadi satu-satunya dasar perbandingan pada penelitian sejenis. Berdasarkan karakteristik arsitektural IndoBERT yang dipra-latih pada korpus berskala masif dan representasi kontekstual bidireksional [13], penelitian ini menghipotesiskan bahwa IndoBERT akan menunjukkan kinerja yang secara statistik signifikan lebih unggul dibandingkan ketiga model pembanding lainnya.

II. METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan rancangan eksperimental komparatif. Variabel bebas dalam penelitian ini adalah empat arsitektur model klasifikasi sentimen Bi-LSTM+FastText, IndoBERT, SVM+TF-IDF, dan Naïve Bayes+TF-IDF sedangkan variabel terikatnya adalah kinerja klasifikasi yang diukur melalui akurasi, presisi, recall, dan weighted F1-score, serta signifikansi statistik perbedaan kinerja antar model yang diukur melalui p-value McNemar's Test dan effect size Cohen's h.

Populasi penelitian adalah seluruh ulasan berbahasa Indonesia pada aplikasi Claude AI di Google Play Store, bersifat tak terbatas karena terus bertambah. Sampel diambil menggunakan teknik purposive sampling dengan kriteria inklusi: ulasan berbahasa Indonesia, mengandung teks bermakna, dan disertai rating valid pada skala 1–5. Data dikumpulkan melalui teknik web scraping menggunakan library google-play-scraper berbasis Python, menghasilkan 2.536 ulasan mentah yang setelah filtrasi awal menyisakan 1.793 ulasan bersih.

Pelabelan data dilakukan secara manual oleh tiga annotator independen berdasarkan Standar Operasional Prosedur (SOP) pelabelan yang mendefinisikan kriteria positif (kepuasan, pujian, rekomendasi), negatif (keluhan, kekecewaan, kritik), dan netral (deskriptif/informatif tanpa muatan emosi dominan), termasuk penanganan kasus sarkasme dan sentimen campuran. Ketiga annotator melabeli jumlah data yang telah disetarakan (1.621 ulasan per annotator) untuk menjamin validitas proses penggabungan label. Validitas inter-annotator agreement diukur menggunakan Cohen's Kappa dengan ambang minimum K>0,6 [14]. Label final ditetapkan melalui majority voting, menghasilkan 1.576 ulasan berlabel.

Pra-pemrosesan teks diterapkan melalui dua skema. Skema umum untuk seluruh model meliputi cleansing, case folding, dan normalisasi kata tidak baku. Skema khusus untuk Bi-LSTM, SVM, dan Naïve Bayes menambahkan stopword removal dan stemming menggunakan PySastrawi, dilanjutkan tokenisasi berbasis kata atau ekstraksi fitur TF-IDF (max_features=5.000, ngram (1,2)). Skema khusus IndoBERT tidak menerapkan stemming/stopword removal untuk menjaga integritas konteks kalimat, menggunakan tokenisasi WordPiece maksimum 128 token. Setelah eliminasi data kosong pasca-preprocessing, dataset final berjumlah 1.566 ulasan, dipartisi menggunakan stratified split 80:20 menjadi 1.252 ulasan data latih dan 314 ulasan data uji — partisi yang identik digunakan pada seluruh model.

Copyright © SENASIF 2026

ISSN: 2598-0076

Model Bi-LSTM+FastText dibangun dengan embedding layer berdimensi 300 yang diinisialisasi bobot pre-trained FastText Bahasa Indonesia (cc.id.300.vec) [10], dilanjutkan lapisan Bidirectional LSTM 64 unit per arah, dropout 0,3, dan lapisan output Softmax; dilatih menggunakan optimizer Adam (learning rate 1e-3) dengan mekanisme early stopping. Model IndoBERT dibangun melalui fine-tuning pada model dasar indobenchmark/indobert-base-p1 (124,4 juta parameter) menggunakan optimizer AdamW (learning rate 2e-5). Model SVM+TF-IDF menggunakan kernel Radial Basis Function dengan parameter C=1,0, sedangkan model Naïve Bayes+TF-IDF menggunakan varian Multinomial dengan parameter alpha=1,0 (Laplace smoothing). Seluruh implementasi dilakukan menggunakan Python 3.10 pada lingkungan Google Colaboratory dengan akselerator GPU NVIDIA T4, memanfaatkan framework PyTorch dan HuggingFace Transformers untuk model deep learning, serta Scikit-learn untuk model baseline.

Evaluasi kinerja menggunakan Confusion Matrix dengan weighted F1-score sebagai metrik utama, dipilih karena kondisi distribusi kelas yang tidak seimbang pada dataset penelitian [15]. Signifikansi perbedaan kinerja antar model diuji menggunakan McNemar's Test pada keenam pasangan model melalui modul scipy.stats.contingency.mcnemar(), dengan ambang signifikansi $p<0,05$, dilengkapi effect size Cohen's h untuk mengukur besaran praktis dari perbedaan yang ditemukan

III. HASIL DAN PEMBAHASAN

Hasil Pelabelan dan Validasi Cohen's Kappa

Ketiga annotator melabeli 1.621 ulasan secara independen. Validitas kesepakatan antar-annotator diukur menggunakan Cohen's Kappa (K), dihitung dengan Persamaan (1).

$$K = (P_o - P_e)/(1 - P_e) \quad (1)$$

dengan P_o adalah proporsi kesepakatan yang diamati dan P_e adalah proporsi kesepakatan yang diharapkan secara kebetulan. Hasil pengukuran disajikan pada Tabel 1.

Tabel 1. Hasil Validasi Cohen's Kappa Antar-Annotator

Pasangan Annotator	Nilai K	Interpretasi
Annotator 1 vs 2	0,8780	Hampir Sempurna
Annotator 1 vs 3	0,9803	Hampir Sempurna
Annotator 2 vs 3	0,8719	Hampir Sempurna
Rata-rata	0,9101	Hampir Sempurna

Rata-rata Cohen's Kappa sebesar 0,9101 tergolong kategori hampir sempurna berdasarkan skala Landis dan Koch [14], jauh melampaui ambang minimum K>0,6. Label final ditentukan melalui majority voting, menghasilkan 1.576 ulasan berlabel dengan distribusi 707 ulasan positif

Copyright © SENASIF 2026
ISSN: 2598-0076

(44,9%), 562 ulasan negatif (35,7%), dan 307 ulasan netral (19,5%). Setelah pra-pemrosesan, dataset final berjumlah 1.566 ulasan (700 positif, 559 negatif, 307 netral), dipartisi menjadi 1.252 ulasan data latih dan 314 ulasan data uji.

Perbandingan Kinerja Empat Model

Tabel 2. Perbandingan Kinerja Empat Model Klasifikasi Sentimen

Model	***Accuracy***	***Weighted F1***	***Macro F1***	***ROC-AUC***
SVM+TF-IDF	0,5446	0,4860	0,4040	0,6124
Naïve Bayes+TF-IDF	0,5414	0,4761	0,3902	0,6073
Bi-LSTM+FastText	0,5318	0,4740	0,3921	0,6067
IndoBERT	0,5223	0,5159	0,4622	0,6263

Tabel 2 menunjukkan pola yang menarik: SVM+TF-IDF mencatat akurasi tertinggi (0,5446), namun IndoBERT unggul pada Weighted F1-Score (0,5159), Macro F1-Score (0,4622), dan ROC-AUC (0,6263). Disosiasi ini terjadi karena akurasi murni tidak mempertimbangkan keseimbangan antar kelas, sehingga model yang cenderung menebak kelas mayoritas dapat mencatat akurasi tinggi meski gagal total pada kelas minoritas netral. Tabel 3 memperjelas fenomena ini melalui perbandingan performa khusus pada kelas netral.

Tabel 3. Perbandingan Performa Kelas Netral Antar Model

Model	***Precision***	***Recall***	***F1-Score***
Naïve Bayes+TF-IDF	0,0000	0,0000	0,0000
Bi-LSTM+FastText	0,0000	0,0000	0,0000
SVM+TF-IDF	0,3333	0,0161	0,0308
IndoBERT	0,2400	0,1935	0,2143

IndoBERT menjadi satu-satunya model yang mengklasifikasikan kelas netral secara bermakna, berhasil mengenali 12 dari 62 ulasan netral pada data uji, sementara Naïve Bayes dan Bi-LSTM+FastText gagal total (F1=0,0000), dan SVM hanya mengenali 1 dari 62 ulasan (F1=0,0308).

Hasil Uji Signifikansi Statistik

Untuk memverifikasi apakah perbedaan kinerja deskriptif bersifat signifikan secara statistik, dilakukan McNemar's Test dan perhitungan effect size Cohen's h pada keenam pasangan model, disajikan pada Tabel 4.

Copyright © SENASIF 2026
ISSN: 2598-0076

Tabel 4. Hasil McNemar's Test dan Effect Size Cohen's h

Model A	Model B	p-value	Signifikan (α=0,05)	Cohen's h
SVM+TF-IDF	Naïve Bayes	1,000	Tidak	0,0064
SVM+TF-IDF	Bi-LSTM+FastText	0,677	Tidak	0,0257
SVM+TF-IDF	IndoBERT	0,494	Tidak	0,0447
Naïve Bayes	Bi-LSTM+FastText	0,783	Tidak	0,0193
Naïve Bayes	IndoBERT	0,556	Tidak	0,0383
Bi-LSTM+FastText	IndoBERT	0,817	Tidak	0,0190

Tabel 4 mengungkap bahwa tidak satu pun dari enam pasangan model menunjukkan perbedaan signifikan secara statistik (seluruh p-value>0,05, berkisar 0,494-1,000). Seluruh nilai effect size Cohen's h tergolong kategori kecil (h<0,05), mengonfirmasi bahwa perbedaan kinerja antar model tidak substansial secara praktis pada ukuran sampel uji yang digunakan. Berdasarkan temuan ini, hipotesis nol tidak dapat ditolak.

Pembahasan

Empat temuan utama dapat didiskusikan dari hasil penelitian ini. Pertama, kinerja moderat keempat model (akurasi 52,2% - 54,5%) mengonfirmasi bahwa ulasan aplikasi AI generatif memiliki kompleksitas semantik yang lebih tinggi dibandingkan domain ulasan aplikasi konvensional yang umumnya mencapai akurasi di atas 80% [9], [16], sejalan dengan sifat penilaian pengguna yang tidak hanya menyasar aspek teknis tetapi juga kecerdasan penalaran model.

Kedua, disosiasi antara akurasi dan Weighted F1-Score pada IndoBERT menjadi bukti empiris penting mengenai urgensi pemilihan metrik evaluasi yang tepat pada kondisi imbalanced data — akurasi semata dapat menyesatkan interpretasi kinerja model. Ketiga, kegagalan Bi-LSTM+FastText melampaui SVM+TF-IDF konvensional, meski telah dilengkapi pre-trained embedding untuk menyetarakan kondisi awal, membuktikan bahwa penyetaraan representasi kata saja tidak cukup mengkompensasi keterbatasan arsitektur RNN dalam menangkap konteks kalimat penuh dibandingkan mekanisme self-attention Transformer [13], sejalan dengan temuan bahwa keunggulan *deep learning* bergantung pada ketersediaan volume data yang memadai [17].

Keempat, kemampuan IndoBERT mengenali kelas netral secara konsisten pada berbagai iterasi pelabelan data yang telah dilakukan dalam penelitian ini meningkat dari F1=0,1707 menjadi F1=0,2143 seiring penyempurnaan proses pelabelan mengindikasikan bahwa keunggulan tersebut bersifat sistematis, bukan kebetulan statistik semata, meski belum terkonfirmasi signifikan secara formal pada uji McNemar's Test. Tidak ditemukannya signifikansi statistik pada seluruh perbandingan model, yang konsisten muncul di berbagai iterasi eksperimen dengan dataset yang sedikit berbeda, menjadi pengingat penting bahwa ukuran sampel uji yang terbatas (n=314)

Copyright © SENASIF 2026
ISSN: 2598-0076

merupakan faktor pembatas utama dalam mendeteksi perbedaan kinerja yang sesungguhnya ada pada populasi.

IV. SIMPULAN

Berdasarkan hasil penelitian mengenai Pengaruh Kualitas Layanan Digital *Mobile* JKN Terhadap Kepuasan Pengguna di Kelurahan Kidul Dalem, dapat disimpulkan bahwa kualitas layanan digital berpengaruh positif dan signifikan terhadap tingkat kepuasan masyarakat. Implemetasi aplikasi ini berhasil meningkatkan efisiensi dan kenyamanan warga dalam mengakses layanan BPJS Kesehatan.

Dimensi utama kualitas layanan—seperti kemudahan penggunaan, keandalan sistem, daya tanggap, dan keamanan data—menjadi faktor kunci pembentuk kepuasan pengguna. Kemudahan pendaftaran antrean online serta kejelasan informasi fitur sangat membantu warga Kelurahan Kidul Dalem menghemat waktu tanpa harus mengantre lama secara fisik di kantor cabang.

DAFTAR PUSTAKA

[1] Anthropic, "Constitutional AI: Harmlessness from AI Feedback," 2022. [Online]. Available: https://arxiv.org/abs/2212.08073.

[2] DataAI, "Global Generative AI Adoption Report," Tech. Rep., 2023.

[3] A. B. C. et al., "Analisis Sentimen Pengguna Aplikasi di Google Play Store," Jurnal Sistem Informasi, vol. 10, no. 2, pp. 45-55, 2022.

[4] D. E. F. dkk., "Tantangan Pelabelan Manual pada Big Data," Jurnal Ilmu Komputer, vol. 8, pp. 11-20, 2021.

[5] G. H. I., "Keterbatasan TF-IDF dalam NLP," Natural Language Engineering, vol. 12, pp. 100-115, 2020.

[6] J. K. L., "Variasi Linguistik dalam Teks Media Sosial Bahasa Indonesia," Jurnal Linguistik Terapan, vol. 5, pp. 20-30, 2019.

[7] M. Illahi et al., "Kombinasi Bi-LSTM dan IndoBERT untuk Data Terbatas," Jurnal AI Indonesia, vol. 3, pp. 40-50, 2023.

[8] Nngrum et al., "Perbandingan BERT dan RNN-LSTM pada Ulasan Perbankan," IEEE Access, vol. 9, pp. 10234-10245, 2022.

[9] R. Mahendra et al., "Analisis Sentimen DeepSeek AI dengan Imbalanced Data," Proceeding of NLP Indonesia, 2024.

[10] E. Grave et al., "Learning Word Vectors for 157 Languages," in LREC, 2018.

[11] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, pp. 153-157, 1947.

[12] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Routledge, 1988.

Copyright © SENASIF 2026
ISSN: 2598-0076

[13] F. A. Purwarianti et al., "IndoBERT: A Pretrained Language Model for Indonesian," in 2020 International Conference on Asian Language Processing (IALP), 2020.

[14] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," Biometrics, vol. 33, no. 1, pp. 159-174, 1977.

[15] S. J. dkk., "Penanganan Imbalanced Data pada Klasifikasi Teks," Jurnal Pembelajaran Mesin, vol. 4, pp. 50-65, 2021.

[16] T. U. V., "Akurasi Model Sentimen pada Ulasan E-Commerce," Jurnal Teknologi Informasi, vol. 11, pp. 88-95, 2022.

[17] W. X. Y., "Deep Learning vs Machine Learning Tradisional: Studi Komparatif," AI Review, vol. 15, pp. 200-215, 2021.

Copyright © SENASIF 2026
ISSN: 2598-0076