Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 25-34
DOI: 10.47065/bulletinds.v6i1.11071
https://ejurnal.seminar-id.com/index.php/bulletinds

Analisis Komparatif Kinerja IndoBERT dan XLM-RoBERTa Dalam Analisis Sentimen Ulasan Google Maps Mitra10 Siliwangi Pamulang

Comparative Analysis of IndoBERT and XLM-RoBERTa Performance in Sentiment Analysis of Google Maps Reviews for Mitra10 Siliwangi Pamulang

M M Rizki Reza Utama*, Sajarwo Anggai, Abu Khalid Rivai

Program Studi Teknik Informatika S-2, Program Pascasarjana, Universitas Pamulang, Banten, Indonesia
Jl. Puspitek, Buaran, Kec. Pamulang, Kota Tangerang Selatan, Banten 15310
Email: [1]rizkireza249@gmail.com, [2]sajarwo@gmail.com, [3]dosen01591@unpam.ac.id
Correspondence Author Email : [1]rizkireza249@gmail.com
Received: **26-08-2026**, Revised: **18-10-2026**, Accepted: **27-10-2026**, Published**: 28-10-2026**

Abstrak−Perkembangan teknologi informasi memungkinkan pelanggan menyampaikan opini melalui platform digital seperti Google Maps. Ulasan pelanggan dapat digunakan untuk memahami persepsi terhadap layanan dan pengalaman berbelanja, tetapi jumlah data yang besar dan bentuk teks yang tidak terstruktur membuat analisis manual menjadi kurang efisien. Penelitian ini bertujuan menganalisis sentimen ulasan pelanggan Mitra10 Siliwangi Pamulang dan membandingkan kinerja IndoBERT dengan XLM-RoBERTa. Data diperoleh melalui web scraping sebanyak 1.305 ulasan yang diklasifikasikan ke dalam sentimen positif, netral, dan negatif, kemudian dibagi menjadi data latih, validasi, dan uji. Evaluasi menggunakan accuracy, weighted precision, weighted recall, weighted F1-score, confusion matrix, dan Exact McNemar Test. XLM-RoBERTa memperoleh accuracy 96,95%, precision 97,25%, recall 96,95%, dan F1-score 97,07%, sedangkan IndoBERT memperoleh accuracy 95,42%, precision 96,90%, recall 95,42%, dan F1-score 95,88%. Secara numerik, XLM-RoBERTa memperoleh accuracy 1,53 poin persentase lebih tinggi. Namun, Exact McNemar Test menghasilkan p-value sebesar 0,625 ($p > 0,05$), sehingga perbedaan kinerja kedua model tidak signifikan secara statistik pada data uji. Hasil tersebut tetap perlu ditafsirkan dengan mempertimbangkan ketidakseimbangan kelas, karena data negatif hanya berjumlah 40 dari1.305ulasan.

Kata Kunci: Analisis Sentimen; Google Maps; Indobert; XLM-Roberta; Transformer; Ulasan Pelanggan

Abstract−Advances in information technology allow customers to share opinions through digital platforms such as Google Maps. Customer reviews can be used to understand perceptions of service and shopping experience, but large volumes of unstructured text make manual analysis less efficient. This study analyzes the sentiment of customer reviews for Mitra10 Siliwangi Pamulang and compares IndoBERT with XLM-RoBERTa. A total of 1,305 reviews were collected through web scraping, classified into positive, neutral, and negative sentiment, and split into training, validation, and test sets. Evaluation used accuracy, weighted precision, weighted recall, weighted F1-score, confusion matrices, and the exact McNemar test. XLM-RoBERTa achieved 96.95% accuracy, 97.25% precision, 96.95% recall, and a 97.07% F1-score, while IndoBERT achieved 95.42% accuracy, 96.90% precision, 95.42% recall, and a 95.88% F1-score. Numerically, XLM-RoBERTa achieved an accuracy that was 1.53 percentage points higher. However, the exact McNemar test produced a p-value of 0.625 ($p > 0.05$), indicating that the performance difference between the two models was not statistically significant on the test set. The results should still be interpreted with the class imbalance in mind, because only 40 of the 1,305 reviews belong to the negative class.

Keywords: Sentiment Analysis; Google Maps; Indobert; XLM-Roberta; Transformer; Customer Reviews

1. PENDAHULUAN

Perkembangan platform digital telah mengubah cara pelanggan menyampaikan evaluasi terhadap produk dan layanan. Ulasan daring tidak lagi hanya berfungsi sebagai testimoni, tetapi juga membentuk sumber data yang dapat dianalisis untuk memahami kepuasan, keluhan, serta persepsi pelanggan. Pada lingkungan ritel, informasi semacam ini relevan karena pengalaman pelanggan dapat mencakup kualitas pelayanan, ketersediaan barang, fasilitas toko, harga, hingga proses transaksi. Penelitian berbasis BERT menunjukkan bahwa ulasan konsumen dapat diklasifikasikan secara efektif ke dalam kategori sentimen ketika representasi kontekstual dimanfaatkan secara tepat [1]. Google Maps menjadi salah satu sumber yang penting karena ulasan diberikan terhadap lokasi fisik dan sering memuat pengalaman pengguna secara langsung.

Mitra10 Siliwangi Pamulang memiliki kumpulan ulasan Google Maps yang terus bertambah dan menggunakan bentuk bahasa yang beragam. Ketika jumlah ulasan meningkat, pembacaan manual menjadi kurang efisien, sulit direproduksi, dan berpotensi menghasilkan interpretasi yang tidak konsisten. Penelitian pada ulasan Google Maps sebelumnya menunjukkan bahwa teknik klasifikasi dapat membantu mengubah teks tidak terstruktur menjadi informasi sentimen yang lebih terukur. Erfina dan Wardani menerapkan Support Vector Machine pada ulasan perguruan tinggi dan memperoleh akurasi tertinggi 85% [2]. Ipmawati dkk. menggunakan SVM pada 2.434 ulasan tempat wisata dan melaporkan akurasi 87% [3]. Fatikhah dkk. menunjukkan bahwa Naive Bayes pada ulasan restoran meningkat dari 86% menjadi 91% setelah penyeimbangan data [4]. Angel dkk. juga menunjukkan bahwa ulasan Google Maps pada layanan

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 25
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 25-34
DOI: 10.47065/bulletinds.v6i1.11071
https://ejurnal.seminar-id.com/index.php/bulletinds

rumah sakit dapat dianalisis menggunakan beberapa algoritma machine learning, dengan Decision Tree mencapai akurasi 92% [5].

Walaupun pendekatan machine learning konvensional memberikan hasil yang baik, representasi berbasis frekuensi kata seperti TF-IDF memiliki keterbatasan ketika suatu kata memiliki makna yang bergantung pada konteks kalimat. Pada ulasan pelanggan berbahasa Indonesia, tantangan ini diperkuat oleh penggunaan singkatan, ejaan tidak baku, bahasa gaul, campuran istilah asing, dan konstruksi negasi. Model Transformer menyediakan representasi kontekstual yang lebih kuat karena setiap token diproses dengan mempertimbangkan keterkaitannya terhadap token lain dalam kalimat. Pada ulasan pelayanan kesehatan berbasis Google Maps, penggunaan BERT/IndoBERT menghasilkan akurasi sekitar 85–86% [6]. Studi ulasan hotel berbasis BERT juga memperoleh akurasi 91,41% setelah fine-tuning [7], sedangkan klasifikasi ulasan iPhone di Amazon menggunakan BERT mencapai accuracy 94,51% dan F1-score 96,92% [8]. Hasil tersebut menguatkan potensi model Transformer untuk data ulasan konsumen.

Untuk bahasa Indonesia, IndoBERT relevan karena prapelatihannya berorientasi pada korpus bahasa Indonesia sehingga representasi tokennya dirancang lebih dekat dengan karakteristik linguistik lokal. Nabiilah dkk. menggunakan model BERT berbasis bahasa Indonesia untuk klasifikasi komentar toksik dan melaporkan performa pengujian yang tinggi [9]. Di sisi lain, XLM-RoBERTa merupakan model multilingual yang memanfaatkan korpus banyak bahasa. Satria dkk. menggunakan XLM-RoBERTa pada klasifikasi sentimen ulasan produk multilingual dan memperoleh accuracy 73% serta weighted F1-score 71% pada skenario yang diteliti [10]. Perbedaan cakupan prapelatihan tersebut menimbulkan pertanyaan empiris: apakah model monolingual bahasa Indonesia selalu lebih tepat untuk ulasan lokal, atau justru model multilingual dapat memperoleh keuntungan dari variasi kata dan istilah yang muncul dalam ulasan Google Maps.

Aspek lain yang penting adalah ketidakseimbangan kelas. Dataset ulasan pelanggan sering didominasi oleh kelas tertentu, sehingga metrik agregat dapat terlihat tinggi walaupun kemampuan model pada kelas minoritas masih terbatas. Wahyuni dkk. menunjukkan bahwa penanganan imbalance pada analisis sentimen dapat meningkatkan kemampuan model mengenali kelas minoritas [11]. Dalam penelitian ini, distribusi asli data dipertahankan dan ketidakseimbangan ditangani menggunakan class weighting pada fungsi loss, bukan melalui oversampling. Pendekatan ini dipilih agar sampel sintetis atau duplikasi data tidak mengubah distribusi observasi asli, tetapi kesalahan pada kelas minoritas tetap diberikan bobot lebih besar selama pelatihan.

State of the art menunjukkan bahwa penelitian sentimen Indonesia telah berkembang dari machine learning konvensional menuju deep learning dan Transformer. Bi-LSTM telah digunakan untuk klasifikasi opini dengan hasil tinggi pada domain tertentu [12], [21], [23], sedangkan studi lain masih mengandalkan kombinasi SVM atau Naive Bayes pada ulasan aplikasi [13], [25] dan data media sosial berskala besar [14], [16], [24], [27], [28]. FastText dan LSTM juga memberikan performa kompetitif pada data media sosial Indonesia [15]. Namun, sebagian besar penelitian tersebut menggunakan Twitter/X, YouTube, aplikasi, hotel, atau objek layanan lain. Kajian yang membandingkan langsung IndoBERT dan XLM-RoBERTa dengan dataset, pembagian data, hyperparameter, dan prosedur evaluasi yang sama pada ulasan Google Maps sektor ritel fisik masih terbatas.

Research gap penelitian ini terletak pada tiga aspek. Pertama, objek penelitian adalah ulasan Google Maps pada ritel fisik, bukan media sosial atau aplikasi. Kedua, dua model Transformer dengan karakter prapelatihan berbeda dibandingkan dalam kondisi eksperimen yang setara, sehingga selisih kinerja tidak dipengaruhi oleh perbedaan hyperparameter utama. Ketiga, perbandingan tidak berhenti pada accuracy dan weighted F1-score, tetapi menggunakan Exact McNemar Test untuk menilai apakah perbedaan prediksi kedua model pada test set yang sama memiliki dukungan statistik. Tujuan penelitian adalah mengklasifikasikan sentimen positif, netral, dan negatif pada ulasan Mitra10 Siliwangi Pamulang, membandingkan IndoBERT dan XLM-RoBERTa, serta menginterpretasikan pola kesalahan kedua model dengan mempertimbangkan ketidakseimbangan kelas. Kontribusi penelitian adalah bukti empiris terukur mengenai performa model monolingual dan multilingual pada domain ulasan ritel berbahasa Indonesia, sekaligus memberikan dasar untuk pengembangan sistem pemantauan sentimen pelanggan yang lebih sistematis.

Literatur yang digunakan menunjukkan bahwa sumber data dan metode klasifikasi pada analisis sentimen sangat beragam. Pada data Google Maps, SVM masih banyak digunakan karena sederhana dan stabil pada representasi TF-IDF [2], [3]. Naive Bayes juga tetap relevan untuk dataset ulasan dengan biaya komputasi yang rendah, tetapi performanya dapat berubah secara signifikan ketika distribusi kelas tidak seimbang [4]. Pendekatan berbasis machine learning semacam ini memberikan baseline yang penting, namun sebagian besar bergantung pada representasi kata yang tidak secara eksplisit memodelkan konteks dua arah. Pada kalimat ulasan yang mengandung negasi, ambiguitas, atau kombinasi pujian dan kritik, representasi kontekstual menjadi semakin penting.

Model deep learning memperluas kemampuan representasi dengan mempelajari pola urutan dan semantik secara otomatis. Bi-LSTM telah menunjukkan akurasi tinggi pada beberapa domain opini [12], sedangkan FastText yang dikombinasikan dengan LSTM juga mampu menghasilkan performa kompetitif pada teks media sosial [15]. Perkembangan berikutnya mengarah pada model Transformer yang memanfaatkan prapelatihan skala besar. BERT pada ulasan hotel [7], ulasan produk [8], dan komentar berbahasa Indonesia [9], [17], [26] menunjukkan bahwa fine-tuning model pralatih dapat menghasilkan kinerja tinggi tanpa membangun representasi bahasa dari awal. Tren tersebut menjadi dasar pemilihan IndoBERT dan XLM-RoBERTa sebagai dua model yang dibandingkan dalam penelitian ini.

Posisi penelitian ini berbeda dari studi terdahulu karena perbandingan difokuskan pada dua model Transformer yang sama-sama kuat tetapi memiliki orientasi prapelatihan berbeda. IndoBERT mewakili model monolingual yang lebih spesifik terhadap bahasa Indonesia, sedangkan XLM-RoBERTa mewakili model multilingual yang secara teori dapat lebih toleran terhadap code-mixing, istilah asing, dan variasi subword. Perbandingan dilakukan dengan dataset,

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 26
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 25-34
DOI: 10.47065/bulletinds.v6i1.11071
https://ejurnal.seminar-id.com/index.php/bulletinds

preprocessing, pembagian data, jumlah epoch, batch size, learning rate, strategi loss, dan test set yang sama. Dengan kontrol eksperimen tersebut, penelitian tidak hanya mencari skor tertinggi, tetapi juga menilai apakah keunggulan numerik benar-benar didukung oleh perbedaan prediksi yang signifikan secara statistik.

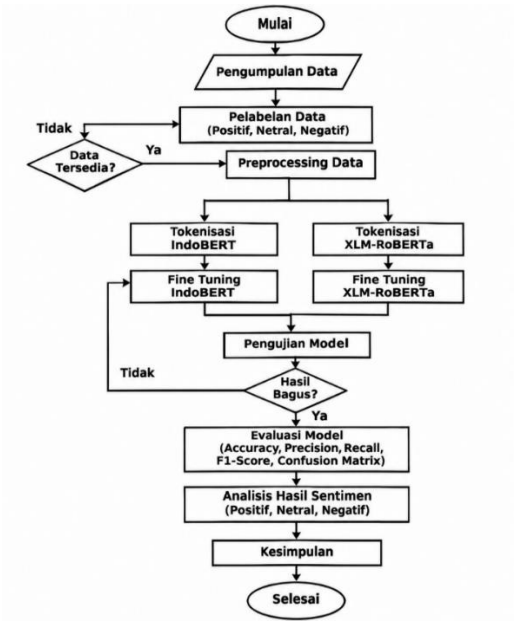
Dari sisi kebaruan, penelitian ini tidak mengklaim menciptakan arsitektur baru. Novelty terletak pada desain komparasi terkontrol pada domain ulasan ritel fisik berbahasa Indonesia dan penggunaan uji statistik berpasangan sebagai pelengkap metrik klasifikasi. Pendekatan ini penting karena selisih satu sampai dua poin persentase sering kali terlihat bermakna secara numerik, padahal pada test set kecil perbedaan tersebut dapat hanya berasal dari beberapa sampel. Dengan Exact McNemar Test, interpretasi hasil menjadi lebih konservatif dan sesuai dengan tujuan penelitian komparatif: membedakan antara model yang sekadar memiliki skor lebih tinggi pada satu eksperimen dengan model yang menunjukkan bukti perbedaan kinerja yang lebih kuat.

Kontribusi penelitian ini mencakup aspek empiris, metodologis, dan praktis. Dari aspek empiris, penelitian ini menyajikan perbandingan kinerja IndoBERT sebagai model monolingual bahasa Indonesia dan XLM-RoBERTa sebagai model multilingual dalam klasifikasi sentimen ulasan Google Maps pada sektor ritel fisik, khususnya Mitra10 Siliwangi Pamulang. Perbandingan tersebut memberikan bukti empiris mengenai kinerja kedua model pada domain ulasan pelanggan berbahasa Indonesia yang memiliki karakteristik berbeda dari data media sosial, aplikasi, maupun ulasan hotel. Dari aspek metodologis, penelitian menerapkan desain eksperimen terkontrol dengan menyamakan pembagian dataset, konfigurasi pelatihan, strategi penanganan ketidakseimbangan kelas, dan prosedur evaluasi untuk meningkatkan keterbandingan hasil. Penelitian ini juga melengkapi evaluasi berbasis metrik klasifikasi dengan Exact McNemar Test untuk menguji signifikansi perbedaan prediksi kedua model pada data uji yang sama, sehingga keunggulan numerik tidak langsung diinterpretasikan sebagai keunggulan statistik. Selain itu, penggunaan class weighting tanpa oversampling mempertahankan distribusi observasi asli sekaligus memberikan perhatian lebih besar terhadap kesalahan klasifikasi pada kelas minoritas. Dari aspek praktis, hasil penelitian dapat menjadi dasar pertimbangan dalam memilih model untuk mengembangkan sistem analisis sentimen yang membantu pengelola ritel memantau persepsi pelanggan secara lebih sistematis. Dengan demikian, kebaruan penelitian ini tidak terletak pada pengembangan arsitektur baru, melainkan pada penerapan komparasi terkontrol antara model monolingual dan multilingual, pengujian statistik berpasangan, serta analisis kinerja klasifikasi dalam konteks ulasan ritel fisik berbahasa Indonesia dengan distribusi kelas yang tidak seimbang.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian menggunakan pendekatan eksperimen komparatif. Seluruh proses dibuat konsisten untuk kedua model agar perbandingan berfokus pada karakteristik IndoBERT dan XLM-RoBERTa, bukan pada perbedaan perlakuan eksperimen. Tahapan penelitian dimulai dari pengumpulan ulasan Google Maps, pelabelan tiga kelas, preprocessing, pembagian dataset, tokenisasi dengan tokenizer masing-masing model, fine-tuning, evaluasi pada data uji yang sama, perbandingan confusion matrix, dan pengujian Exact McNemar. Alur ringkas penelitian ditunjukkan pada Gambar 1.



Gambar 1. Tahapan penelitian komparatif IndoBERT dan XLM-RoBERTa

Pada gambar 1 menjelaskan tahapan penelitian analisis sentimen untuk membandingkan kinerja model IndoBERT dan XLM-RoBERTa. Penelitian diawali dengan pengumpulan data ulasan, kemudian dilanjutkan dengan pelabelan ke

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 27
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 25-34
DOI: 10.47065/bulletinds.v6i1.11071
https://ejurnal.seminar-id.com/index.php/bulletinds

dalam tiga kategori, yaitu positif, netral, dan negatif. Data yang telah terlabeli selanjutnya melalui tahap preprocessing untuk membersihkan dan mempersiapkan teks sebelum digunakan dalam proses pemodelan.

Tahap berikutnya adalah tokenisasi menggunakan tokenizer masing-masing model, kemudian dilanjutkan dengan proses fine-tuning untuk menyesuaikan model dengan dataset penelitian. Kedua model kemudian diuji untuk mengetahui kemampuan klasifikasinya. Apabila hasil pengujian belum memenuhi kriteria yang ditetapkan, proses penyesuaian dan pengujian model dapat diulangi hingga diperoleh hasil yang sesuai.

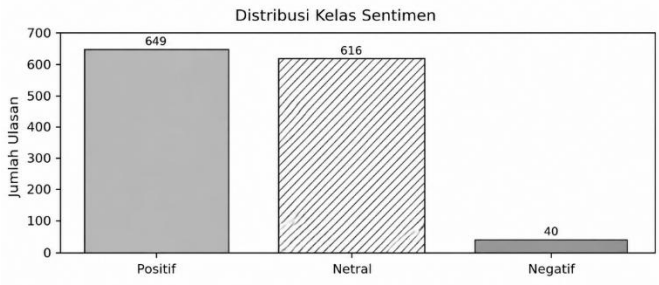
Selanjutnya, kinerja kedua model dievaluasi menggunakan metrik accuracy, precision, recall, F1-score, dan confusion matrix. Hasil evaluasi digunakan untuk membandingkan kemampuan IndoBERT dan XLM-RoBERTa dalam mengklasifikasikan sentimen serta menganalisis hasil prediksi pada setiap kategori. Tahap akhir penelitian adalah menyusun kesimpulan berdasarkan hasil evaluasi dan analisis komparatif untuk menentukan model dengan kinerja terbaik dalam analisis sentimen ulasan.

2.2 Dataset dan Preprocessing

Data berasal dari ulasan Google Maps Mitra10 Siliwangi Pamulang pada periode Oktober 2022 sampai April 2026. Setelah seleksi dan pemeriksaan kualitas, dataset akhir terdiri dari 1.305 ulasan. Teks ulasan digunakan sebagai masukan utama klasifikasi, sedangkan informasi lain seperti rating hanya berfungsi sebagai atribut pendukung. Data dibagi menjadi 1.043 data latih, 131 data validasi, dan 131 data uji. Pembagian terpisah ini digunakan agar model akhir dievaluasi pada sampel yang tidak digunakan dalam pelatihan.

Preprocessing meliputi penghapusan data duplikat, case folding, dan normalisasi format teks. Duplikasi diperiksa berdasarkan kombinasi reviewer, tanggal ulasan, teks ulasan, dan nama tempat. Case folding mengubah karakter alfabet menjadi huruf kecil. Normalisasi tidak melakukan substitusi leksikal berbasis kamus; proses difokuskan pada penyatuan spasi berlebih, tab, dan newline menjadi satu spasi serta penghapusan spasi di awal dan akhir teks. Pendekatan ini mempertahankan kata informal dan variasi ejaan agar perbedaan kemampuan tokenizer kedua model tetap dapat diamati secara alami.

Distribusi kelas terdiri atas 649 ulasan positif, 616 netral, dan 40 negatif. Kelas negatif hanya mencakup sekitar 3,07% dataset, sehingga ketidakseimbangan kelas cukup tinggi. Penelitian tidak menggunakan SMOTE atau oversampling. Sebagai gantinya, class weighting dihitung dari distribusi label data latih dan diterapkan pada weighted CrossEntropyLoss. Dengan cara ini, kesalahan pada kelas minoritas berkontribusi lebih besar terhadap loss tanpa membuat duplikasi sampel. Distribusi kelas ditunjukkan pada Gambar 2.



Gambar 2. Distribusi kelas sentimen pada dataset

2.3 Tokenisasi, Fine-Tuning, dan Evaluasi

IndoBERT menggunakan tokenizer dari model pralatih indobenchmark/indobert-base-p1. XLM-RoBERTa menggunakan tokenizer berbasis SentencePiece. Teks diubah menjadi input_ids dan attention_mask, kemudian dibatasi maksimum 128 token melalui truncation. Padding dilakukan secara dinamis menggunakan DataCollatorWithPadding sampai panjang sequence terpanjang pada setiap batch. Penggunaan padding dinamis mengurangi komputasi pada ulasan pendek tanpa mengubah batas maksimum input.

Kedua model di-fine-tuning dengan konfigurasi yang sama: 5 epoch, batch size 8, learning rate 2×10^{-5}, optimizer AdamW, weight decay 0,01, linear learning-rate scheduler tanpa warmup, serta random seed 42. Evaluasi dan checkpoint dilakukan setiap epoch. Model akhir dipilih berdasarkan weighted F1-score terbaik pada data validasi. Konfigurasi yang sama dimaksudkan untuk mengendalikan variabel eksperimen sehingga perbedaan hasil lebih dapat dikaitkan dengan model pralatih dan tokenizer yang digunakan.

Kinerja model dihitung menggunakan accuracy, weighted precision, weighted recall, weighted F1-score, dan confusion matrix. Weighted average dipilih karena distribusi kelas tidak seimbang sehingga kontribusi setiap kelas dipertimbangkan sesuai jumlah sampelnya. Confusion matrix dianalisis untuk melihat pola kesalahan pada masing-masing kelas. Selanjutnya, Exact McNemar Test diterapkan pada pasangan prediksi IndoBERT dan XLM-RoBERTa pada 131 data uji yang sama. Hipotesis nol menyatakan tidak terdapat perbedaan kinerja klasifikasi yang signifikan; H0 ditolak apabila p-value lebih kecil dari 0,05.

Konfigurasi parameter yang digunakan dalam proses fine-tuning model IndoBERT dan XLM-RoBERTa disajikan pada Tabel 1. Kedua model menggunakan konfigurasi parameter yang sama untuk memastikan proses pelatihan dan evaluasi dilakukan secara konsisten.

Tabel 1. Konfigurasi utama fine-tuning

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 28
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 25-34
DOI: 10.47065/bulletinds.v6i1.11071
https://ejurnal.seminar-id.com/index.php/bulletinds

Parameter	Nilai	Keterangan
Epoch	5	Sama untuk kedua model
Batch size	8	Train dan evaluasi
Learning rate	2×10^{-5}	Fine-tuning
Max sequence	128 token	Truncation
Optimizer	AdamW	Weight decay 0,01
Loss	Weighted CrossEntropyLoss	Class weighting
Random seed	42	TrainingArguments

2.4 Kontrol Eksperimen dan Reprodusibilitas

Eksperimen dirancang agar kedua model menerima perlakuan yang setara. Dataset tidak diubah di antara proses fine-tuning IndoBERT dan XLM-RoBERTa, sehingga setiap model menggunakan pembagian train, validation, dan test yang identik. Hyperparameter utama juga disamakan, termasuk jumlah epoch, batch size, learning rate, maximum sequence length, weight decay, strategi evaluasi per epoch, pemilihan checkpoint berdasarkan weighted F1-score validasi, serta random seed 42. Pengendalian ini diperlukan karena perubahan kecil pada split atau konfigurasi training dapat menghasilkan selisih performa yang sulit dibedakan dari perbedaan karakteristik model.

Walaupun kontrol eksperimen telah diterapkan, penelitian hanya menggunakan satu random seed dan satu konfigurasi hyperparameter. Oleh karena itu, hasil yang diperoleh merepresentasikan satu realisasi eksperimen yang terkontrol, bukan distribusi performa dari banyak pengulangan. Untuk menjaga validitas interpretasi, model dengan skor lebih tinggi hanya disebut unggul secara numerik pada dataset penelitian. Klaim superioritas umum tidak dibuat tanpa pengulangan multi-seed, hyperparameter tuning, dan test set yang lebih besar. Prinsip ini juga menjadi alasan penggunaan Exact McNemar Test pada prediksi data uji yang sama.

3. HASIL DAN PEMBAHASAN

3.1 Karakteristik Dataset dan Dampak Ketidakseimbangan Kelas

Dataset menunjukkan dua kelas mayoritas yang relatif seimbang, yaitu positif 649 data dan netral 616 data, sedangkan kelas negatif hanya 40 data. Komposisi ini memperlihatkan bahwa sebagian besar pelanggan memberikan ulasan positif atau deskriptif netral, sementara ulasan negatif relatif sedikit. Dari sisi operasional, dominasi positif dapat menjadi indikator pengalaman berbelanja yang cenderung baik. Namun, dari sisi pemodelan, proporsi negatif yang kecil menimbulkan risiko model kurang memperoleh variasi pola bahasa negatif selama pelatihan.

Konsekuensi ketidakseimbangan terlihat lebih kuat pada data uji. Dari 131 sampel uji, hanya empat sampel termasuk kelas negatif. Artinya, satu kesalahan prediksi pada kelas tersebut dapat mengubah recall negatif sebesar 25 poin persentase. Oleh karena itu, keberhasilan suatu model mengenali 4/4 atau 3/4 sampel negatif tidak dapat langsung digeneralisasikan sebagai kemampuan yang stabil pada seluruh bentuk keluhan pelanggan. Pertimbangan ini penting karena model klasifikasi yang digunakan untuk pemantauan layanan justru perlu sensitif terhadap keluhan, walaupun jumlahnya lebih sedikit. Penanganan imbalance menggunakan class weighting mengurangi dominasi kelas mayoritas pada loss, tetapi tidak menghilangkan keterbatasan statistik akibat sedikitnya contoh negatif.

Temuan ini konsisten dengan literatur yang menunjukkan bahwa distribusi kelas dapat memengaruhi kualitas klasifikasi. Wahyuni dkk. melaporkan peningkatan kinerja ketika kelas minoritas ditangani secara eksplisit [11]. Pada ulasan Google Maps restoran, Fatikhah dkk. juga menunjukkan bahwa penyeimbangan data meningkatkan akurasi Naive Bayes [4]. Penelitian ini memilih class weighting alih-alih oversampling agar perbandingan IndoBERT dan XLM-RoBERTa dilakukan pada distribusi data asli. Implikasinya, nilai agregat utama tetap perlu dibaca bersama confusion matrix dan jumlah sampel per kelas.

3.2 Hasil Model IndoBERT

IndoBERT memperoleh accuracy 95,42%, weighted precision 96,90%, weighted recall 95,42%, dan weighted F1-score 95,88%. Dari 131 data uji, sebanyak 125 data diklasifikasikan dengan benar dan enam data salah. Confusion matrix menunjukkan bahwa empat data negatif seluruhnya diprediksi benar. Pada kelas netral, 60 data diprediksi benar dan dua data salah diprediksi sebagai positif. Pada kelas positif, 61 data diprediksi benar dan empat data salah diprediksi sebagai negatif.

Hasil tersebut menunjukkan bahwa IndoBERT mampu memodelkan konteks bahasa Indonesia pada ulasan ritel dengan sangat baik. Nilai weighted precision yang lebih tinggi daripada accuracy menunjukkan bahwa ketika model menetapkan suatu kelas, prediksi tersebut secara umum memiliki ketepatan tinggi. Kesalahan terbesar muncul pada empat ulasan positif yang diprediksi negatif. Pola ini dapat terjadi ketika kalimat positif memuat token yang secara lokal memiliki nuansa negatif, misalnya struktur negasi seperti “tidak mengecewakan” atau komentar yang menggabungkan pujian dan kritik ringan. Karena model memproses konteks secara bidirectional, sebagian besar kasus dapat dikenali, tetapi sampel yang ambigu tetap menimbulkan kesalahan.

Kinerja IndoBERT pada penelitian ini lebih tinggi dibandingkan beberapa penelitian BERT pada domain Indonesia yang dilaporkan sebelumnya. Ardiansyah dkk. memperoleh akurasi sekitar 85–86% pada ulasan pelayanan kesehatan Google Maps [6], sedangkan Chandradev dkk. melaporkan akurasi 91,41% pada review hotel [7]. Perbedaan angka tidak dapat dianggap sebagai bukti superioritas langsung karena objek, distribusi kelas, ukuran dataset, prosedur

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 29
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 25-34
DOI: 10.47065/bulletinds.v6i1.11071
https://ejurnal.seminar-id.com/index.php/bulletinds

pelabelan, dan skema evaluasi berbeda. Namun, hasil tersebut memperlihatkan bahwa fine-tuning model pralatih bahasa Indonesia dapat bekerja efektif pada ulasan ritel ketika pipeline preprocessing dan pelatihan dijaga konsisten.

Dibandingkan pendekatan non-Transformer, accuracy IndoBERT juga lebih tinggi daripada hasil SVM 85% pada ulasan perguruan tinggi [2] dan 87% pada ulasan wisata [3]. Perbandingan ini bersifat kontekstual, bukan eksperimen head-to-head, sehingga tidak digunakan untuk menyimpulkan Transformer selalu lebih unggul. Keuntungan utama IndoBERT dalam penelitian ini adalah kemampuannya memanfaatkan representasi kontekstual tanpa harus membangun fitur TF-IDF secara manual.

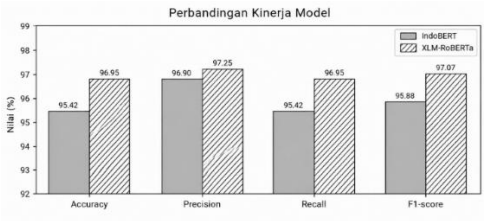
3.3 Hasil Model XLM-RoBERTa

XLM-RoBERTa memperoleh accuracy 96,95%, weighted precision 97,25%, weighted recall 96,95%, dan weighted F1-score 97,07%. Model melakukan empat kesalahan dari 131 data uji dan mengklasifikasikan 127 data dengan benar. Pada kelas negatif, tiga dari empat sampel diprediksi benar dan satu diprediksi positif. Pada kelas netral, 61 data diprediksi benar dan satu diprediksi positif. Pada kelas positif, 63 data diprediksi benar dan dua diprediksi negatif.

Secara agregat, XLM-RoBERTa menghasilkan nilai tertinggi pada seluruh metrik utama. Salah satu penjelasan yang masuk akal adalah kapasitas multilingual dan tokenizer SentencePiece yang dapat memproses variasi subword lintas bahasa. Ulasan Google Maps sering memuat istilah asing, singkatan, nama produk, dan variasi ejaan. Karakteristik tersebut dapat memberi keuntungan bagi representasi multilingual. Namun, penelitian ini tidak melakukan analisis ablation terhadap jenis token atau proporsi code-mixing, sehingga hubungan antara kemampuan multilingual dan peningkatan kinerja harus diposisikan sebagai interpretasi, bukan kesimpulan kausal.

Hasil penelitian Satria dkk. menunjukkan bahwa XLM-RoBERTa dapat digunakan pada ulasan produk multilingual [10], tetapi akurasi yang mereka peroleh berbeda karena tugas dan konfigurasi penelitian tidak sama. Pada penelitian ini, XLM-RoBERTa memperoleh hasil tinggi karena diuji pada dataset tiga kelas dengan domain dan distribusi tertentu. Karena test set hanya 131 sampel, kenaikan dua prediksi benar dibandingkan IndoBERT cukup untuk menghasilkan selisih accuracy sekitar 1,53 poin persentase. Oleh sebab itu, keunggulan numerik perlu diuji secara statistik sebelum dinyatakan sebagai perbedaan kinerja yang bermakna.

Perbandingan kinerja IndoBERT dan XLM-RoBERTa berdasarkan nilai accuracy, precision, recall, dan F1-score disajikan pada Gambar 3. Hasil menunjukkan bahwa XLM-RoBERTa memperoleh nilai lebih tinggi pada seluruh metrik dibandingkan IndoBERT. Selisih accuracy kedua model sebesar 1,53 poin persentase. Namun, keunggulan numerik tersebut perlu ditinjau bersama hasil pengujian statistik untuk mengetahui apakah perbedaan kinerja kedua model signifikan secara statistik.



Gambar 3. Perbandingan metrik agregat IndoBERT dan XLM-RoBERTa

Hasil evaluasi kinerja model IndoBERT dan XLM-RoBERTa pada data uji berdasarkan metrik accuracy, precision, recall, dan F1-score disajikan pada Tabel 2. Perbandingan ini bertujuan untuk mengetahui kemampuan kedua model dalam melakukan klasifikasi sentimen berdasarkan nilai metrik evaluasi yang diperoleh.

Tabel 2. Perbandingan kinerja model pada data uji

Model	Accuracy	Precision	Recall	F1-score
IndoBERT	95,42%	96,90%	95,42%	95,88%
XLM-RoBERTa	96,95%	97,25%	96,95%	97,07%

Berdasarkan Tabel 2, model XLM-RoBERTa memperoleh nilai accuracy sebesar 96,95%, precision sebesar 97,25%, recall sebesar 96,95%, dan F1-score sebesar 97,07%. Sementara itu, IndoBERT memperoleh nilai accuracy sebesar 95,42%, precision sebesar 96,90%, recall sebesar 95,42%, dan F1-score sebesar 95,88%. Hasil tersebut menunjukkan bahwa XLM-RoBERTa memiliki kinerja lebih baik dibandingkan IndoBERT pada seluruh metrik evaluasi yang digunakan dalam penelitian ini. Dengan demikian, XLM-RoBERTa menunjukkan kemampuan klasifikasi sentimen yang lebih baik berdasarkan hasil pengujian pada dataset penelitian.

3.4 Perbandingan Confusion Matrix dan Exact McNemar Test

Perbedaan kedua model tidak hanya terlihat pada jumlah prediksi benar, tetapi juga pada pola kesalahan. IndoBERT melakukan enam kesalahan, sedangkan XLM-RoBERTa melakukan empat kesalahan. IndoBERT lebih baik pada kelas negatif karena mengenali seluruh empat sampel, sementara XLM-RoBERTa salah pada satu sampel negatif. Sebaliknya, XLM-RoBERTa lebih baik pada kelas positif dan netral. Kondisi ini penting untuk konteks bisnis. Jika prioritas utama adalah tidak melewatkan keluhan pelanggan, kesalahan negatif menjadi positif pada XLM-RoBERTa perlu diperhatikan, walaupun akurasi agregatnya lebih tinggi.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 30
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 25-34
DOI: 10.47065/bulletinds.v6i1.11071
https://ejurnal.seminar-id.com/index.php/bulletinds

Perbandingan prediksi berpasangan menghasilkan 124 sampel yang benar oleh kedua model, tiga sampel yang salah oleh kedua model, satu sampel yang benar hanya oleh IndoBERT, serta tiga sampel yang benar hanya oleh XLM-RoBERTa. Dengan demikian, pasangan discordant terdiri atas b=1 dan c=3. Exact McNemar Test menghasilkan p-value 0,625 pada tingkat signifikansi 0,05. Karena p-value lebih besar dari 0,05, hipotesis nol tidak ditolak. Hasil ini berarti data uji tidak menyediakan bukti statistik yang cukup untuk menyatakan bahwa perbedaan kinerja kedua model signifikan.

Interpretasi tersebut membatasi klaim penelitian. XLM-RoBERTa dapat dinyatakan memiliki hasil numerik lebih tinggi pada dataset ini, tetapi tidak dapat dinyatakan secara statistik lebih unggul daripada IndoBERT. Jumlah pasangan discordant hanya empat, sehingga daya uji statistik juga terbatas. Dengan sampel uji yang lebih besar, pola perbedaan dapat berubah. Oleh karena itu, penelitian lanjutan perlu menggunakan test set yang lebih besar, beberapa random seed, dan replikasi pelatihan untuk menilai kestabilan hasil.

Untuk menguji signifikansi perbedaan kinerja klasifikasi antara IndoBERT dan XLM-RoBERTa pada data uji yang sama, penelitian ini menggunakan Exact McNemar Test. Pengujian dilakukan berdasarkan jumlah prediksi benar dan salah dari kedua model pada 131 data uji. Hasil perbandingan prediksi dan pengujian statistik tersebut disajikan pada Tabel 3.

Tabel 3. Pasangan prediksi untuk Exact McNemar Test (p-value = 0,625)

IndoBERT / XLM-R	Benar	Salah	Total
Benar	124	1	125
Salah	3	3	6
Total	127	4	131

Berdasarkan Tabel 4, terdapat 124 data yang diprediksi benar oleh kedua model, satu data yang diprediksi benar oleh IndoBERT tetapi salah oleh XLM-RoBERTa, tiga data yang diprediksi salah oleh IndoBERT tetapi benar oleh XLM-RoBERTa, serta tiga data yang diprediksi salah oleh kedua model. Hasil Exact McNemar Test menghasilkan p-value sebesar 0,625, lebih besar daripada taraf signifikansi 0,05. Dengan demikian, tidak terdapat bukti statistik yang cukup untuk menyatakan adanya perbedaan kinerja klasifikasi yang signifikan antara kedua model pada dataset pengujian ini. Meskipun XLM-RoBERTa memperoleh nilai metrik evaluasi agregat yang lebih tinggi, keunggulan tersebut belum terbukti signifikan secara statistik berdasarkan pengujian ini.

3.5 Pembahasan

Temuan penelitian ini menunjukkan bahwa XLM-RoBERTa memperoleh kinerja agregat lebih tinggi daripada IndoBERT, dengan accuracy 96,95% dan weighted F1-score 97,07%, sedangkan IndoBERT memperoleh accuracy 95,42% dan weighted F1-score 95,88%. Hasil ini sejalan dengan penelitian Setyawan dkk. [8] yang melaporkan bahwa model BERT mampu mencapai accuracy 94,51% dan F1-score 96,92% pada ulasan produk iPhone di Amazon. Meskipun hasil penelitian ini menunjukkan nilai yang lebih tinggi pada accuracy, perbandingan tersebut bersifat kontekstual karena perbedaan bahasa, domain, jumlah data, distribusi kelas, dan prosedur evaluasi.

Jika dibandingkan dengan penelitian pada ulasan Google Maps, hasil kedua model pada penelitian ini juga lebih tinggi daripada akurasi yang dilaporkan oleh Erfina dan Wardani [2], yaitu 85% menggunakan SVM pada ulasan perguruan tinggi, serta Ipmawati dkk. [3], yaitu 87% menggunakan SVM pada ulasan tempat wisata. Ardiansyah dkk. [6] melaporkan akurasi sekitar 85–86% dengan BERT pada ulasan layanan kesehatan Google Maps, sedangkan Angel dkk. [5] memperoleh akurasi 92% menggunakan Decision Tree pada ulasan rumah sakit. Perbedaan ini memperlihatkan bahwa model Transformer memberikan hasil kompetitif pada dataset Mitra10, tetapi tidak dapat dijadikan bukti bahwa Transformer selalu lebih unggul karena eksperimen tidak dilakukan pada dataset dan kondisi pengujian yang sama.

Pada penelitian terdahulu, Fatikhah dkk. [4] melaporkan peningkatan akurasi Naive Bayes dari 86% menjadi 91% setelah penyeimbangan data. Penelitian ini menggunakan pendekatan berbeda, yaitu class weighting tanpa oversampling, untuk mempertahankan distribusi ulasan asli. Walaupun nilai metrik agregat kedua model tinggi, kelas negatif hanya terdiri atas 40 dari 1.305 ulasan dan empat sampel pada data uji. Oleh karena itu, hasil perlu dibaca bersama confusion matrix: IndoBERT mengenali seluruh empat sampel negatif, sedangkan XLM-RoBERTa mengenali tiga sampel negatif. Temuan ini menunjukkan bahwa kinerja agregat yang lebih tinggi tidak selalu berarti lebih baik pada setiap kelas.

Untuk model multilingual, Satria dkk. [10] menunjukkan penerapan XLM-RoBERTa pada klasifikasi sentimen ulasan produk multilingual, dengan hasil yang bergantung pada skenario tugas yang digunakan. Pada penelitian ini, XLM-RoBERTa memperoleh nilai agregat tertinggi, tetapi Exact McNemar Test menghasilkan p-value 0,625 ($p > 0,05$). Dengan demikian, selisih accuracy sebesar 1,53 poin persentase belum memberikan bukti statistik yang cukup untuk menyatakan bahwa kinerja kedua model berbeda secara signifikan. Secara keseluruhan, hasil penelitian mendukung bahwa IndoBERT dan XLM-RoBERTa sama-sama kompetitif untuk analisis sentimen ulasan ritel berbahasa Indonesia. Pemilihan model sebaiknya mempertimbangkan tujuan penggunaan, terutama apabila deteksi ulasan negatif menjadi prioritas, serta perlu divalidasi kembali menggunakan data uji yang lebih besar.

4. KESIMPULAN

Penelitian ini berhasil membandingkan IndoBERT dan XLM-RoBERTa untuk klasifikasi tiga kelas sentimen pada 1.305 ulasan Google Maps Mitra10 Siliwangi Pamulang menggunakan pipeline eksperimen yang konsisten. Dataset dibagi

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 31
This work is licensed under a Creative Commons Attribution 4.0 International License.

ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 25-34
DOI: 10.47065/bulletinds.v6i1.11071
https://ejurnal.seminar-id.com/index.php/bulletinds

menjadi 1.043 data latih, 131 data validasi, dan 131 data uji, dengan distribusi 649 ulasan positif, 616 netral, dan 40 negatif. Ketidakseimbangan kelas ditangani melalui class weighting pada weighted CrossEntropyLoss tanpa oversampling. IndoBERT memperoleh accuracy 95,42%, weighted precision 96,90%, weighted recall 95,42%, dan weighted F1-score 95,88%. XLM-RoBERTa memperoleh accuracy 96,95%, weighted precision 97,25%, weighted recall 96,95%, dan weighted F1-score 97,07%, sehingga secara numerik unggul 1,53 poin persentase pada accuracy. Namun, Exact McNemar Test menghasilkan p-value 0,625 pada α=0,05, sehingga tidak terdapat bukti yang cukup untuk menyatakan perbedaan kinerja kedua model signifikan secara statistik pada test set yang digunakan. Analisis confusion matrix juga menunjukkan trade-off: IndoBERT mengenali seluruh empat sampel negatif, sedangkan XLM-RoBERTa membuat lebih sedikit kesalahan secara total. Karena kelas negatif hanya 3,07% dari keseluruhan data dan hanya empat sampel terdapat pada data uji, hasil per kelas negatif harus ditafsirkan secara hati-hati. Kontribusi utama penelitian adalah menunjukkan bahwa model monolingual dan multilingual sama-sama kompetitif pada ulasan ritel berbahasa Indonesia, serta bahwa selisih metrik agregat yang kecil perlu dilengkapi pengujian statistik berpasangan sebelum ditafsirkan sebagai superioritas model. Keterbatasan penelitian mencakup objek yang hanya satu lokasi ritel, ukuran kelas negatif yang kecil, satu konfigurasi hyperparameter, satu random seed, dan belum adanya evaluasi efisiensi komputasi. Penelitian selanjutnya disarankan memperluas dataset ke beberapa cabang atau objek ritel, meningkatkan representasi kelas negatif, mengulang fine-tuning dengan beberapa random seed, melakukan hyperparameter tuning sistematis, serta mengukur waktu training, waktu inferensi, konsumsi memori, dan jumlah parameter agar pemilihan model dapat mempertimbangkan akurasi sekaligus biaya implementasi.

5. PERNYATAAN PENULIS

5.1 Kontribusi Penulis

Konseptualisasi: M.M. Rizki Reza Utama, Sajarwo Anggai, dan Abu Khalid Rivai;
Metodologi: M.M. Rizki Reza Utama, Sajarwo Anggai, dan Abu Khalid Rivai;
Perangkat Lunak: M.M. Rizki Reza Utama; Validasi: M.M. Rizki Reza Utama, Sajarwo Anggai, dan Abu Khalid Rivai;
Analisis Formal: M.M. Rizki Reza Utama;
Investigasi dan Kurasi Data: M.M. Rizki Reza Utama;
Penulisan Draf Awal: M.M. Rizki Reza Utama;
Peninjauan dan Penyuntingan: M.M. Rizki Reza Utama, Sajarwo Anggai, dan Abu Khalid Rivai;
Visualisasi: M.M. Rizki Reza Utama.

5.2 Ketersediaan Data

Data penelitian tersedia berdasarkan permintaan yang wajar kepada penulis korespondensi dan dengan memperhatikan ketentuan penggunaan data dari platform sumber.

5.3 Pendanaan

Penelitian ini didanai secara mandiri oleh penulis dan tidak menerima pendanaan eksternal dari institusi, lembaga pemerintah, organisasi, maupun pihak swasta.

5.4 Pernyataan Dewan Peninjau Institusional

Tidak berlaku.

5.5 Pernyataan Persetujuan Berdasarkan Informasi

Tidak berlaku.

5.6 Pernyataan Konflik Kepentingan

Para penulis menyatakan tidak memiliki konflik kepentingan yang diketahui terkait pekerjaan yang dilaporkan dalam artikel ini. Pernyataan ini perlu dikonfirmasi seluruh penulis sebelum pengiriman.

5.7 Penggunaan AI Generatif dan Teknologi Berbantuan AI dalam Proses Penulisan

AI generatif digunakan untuk membantu peringkasan, restrukturisasi, dan peningkatan kejelasan bahasa pada manuskrip. Seluruh isi, data, interpretasi, sitasi, dan versi akhir harus ditinjau serta dipertanggungjawabkan oleh para penulis.

REFERENCES

[1] N. Al Hafidh and A. Al-Karawi , “Advanced sentiment analysis of Amazon electronics reviews leveraging BERT: Model optimization and evaluation,” *Procedia Computer Science*, vol. 258, pp. 3608–3618, 2025, https://doi.org/10.1016/j.procs.2025.04.616

[2] A. Erfina and N. R. Wardani, “Analisis sentimen perguruan tinggi termewah di Indonesia menurut ulasan Google Maps menggunakan Support Vector Machine (SVM),” *Jurnal Manajemen Informatika dan Sistem Informasi*, vol. 5, no. 1, pp. 77–85, 2022, https://doi.org/10.36595/misi.v5i1.591

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 32
This work is licensed under a Creative Commons Attribution 4.0 International License.

[3] J. Ipmawati, Saifulloh, and Kusnawi, "Analisis sentimen tempat wisata berdasarkan ulasan pada Google Maps menggunakan algoritma Support Vector Machine," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 1, pp. 247–256, 2024, https://doi.org/10.57152/malcom.v4i1.1066
[4] T. Z. Fatikhah, M. Muhtarom, and I. Oktaviani, "Analisis sentimen ulasan Mie Gacoan Solo Veteran di Google Maps menggunakan algoritma Naive Bayes," *ZONAsi: Jurnal Sistem Informasi*, vol. 6, no. 3, pp. 649–658, 2024, https://doi.org/10.31849/zn.v6i3.21194
[5] A. C. T. Angel, V. H. Pranatawijaya, and Widiatry, "Analisis sentimen dan emosi dari ulasan Google Maps untuk layanan rumah sakit di Palangka Raya menggunakan machine learning," *KONSTELASI*, vol. 4, no. 1, pp. 35–49, 2024, https://doi.org/10.24002/konstelasi.v4i1.8924
[6] Ardiansyah, A. S. Widagdo, K. N. Qodri, F. E. N. Saputro, and N. A. Rizky P., "Analisis sentimen terhadap pelayanan kesehatan berdasarkan ulasan Google Maps menggunakan BERT," *Jurnal Fasilkom*, vol. 13, no. 2, pp. 326–333, 2023, https://doi.org/10.37859/jf.v13i02.5170
[7] V. Chandradev, I. M. A. D. Suarjaya, and I. P. A. Bayupati, "Analisis sentimen review hotel menggunakan metode deep learning BERT," *Jurnal Buana Informatika*, vol. 14, no. 2, pp. 107–116, 2023, https://doi.org/10.24002/jbi.v14i02.7244
[8] A. F. Setyawan, A. D. P. Ariyanto, F. K. Fikriah, and R. I. Nugraha, "Analisis sentimen ulasan iPhone di Amazon menggunakan model deep learning BERT berbasis transformer," *Elkom: Jurnal Elektronika dan Komputer*, vol. 17, no. 2, pp. 447–452, 2024, https://doi.org/10.51903/elkom.v17i2.2150
[9] G. Z. Nabiilah, S. Y. Prasetyo, Z. N. Izdihar, and A. S. Girsang, "BERT base model for toxic comment analysis on Indonesian social media," *Procedia Computer Science*, vol. 216, pp. 714–721, 2023, https://doi.org/10.1016/j.procs.2022.12.188
[10] G. P. Satria, K. S. Gunawan, and Z. E. Rasjid, "Aspect-based sentiment analysis of multilingual customer product reviews: Leveraging unsupervised aspect extraction with ABAE for XLM-RoBERTa-based classification," *Procedia Computer Science*, vol. 269, pp. 239–248, 2025, https://doi.org/10.1016/j.procs.2025.08.276
[11] N. A. Wahyuni, D. P. Ayu, and H. Irsyad, "Analisis sentimen di YouTube terhadap kenaikan UKT menggunakan metode Support Vector Machine," *Arcitech: Journal of Computer Science and Artificial Intelligence*, vol. 4, no. 1, pp. 57–71, 2024, https://doi.org/10.29240/arcitech.v4i1.10829
[12] F. Bouchra, I. M. A. D. Suarjaya, and N. K. D. Rusjayanthi, "Analisis sentimen masyarakat terhadap tayangan televisi nasional menggunakan metode deep learning," *Jurnal Buana Informatika*, vol. 15, no. 2, pp. 89–99, 2024, https://doi.org/10.24002/jbi.v15i2.9118
[13] A. Y. Lubis and M. Y. H. Setyawan, "Analisis sentimen terhadap aplikasi Pospay menggunakan algoritma Support Vector Machine dan Naive Bayes," *Jurnal Teknologi dan Sistem Informasi Bisnis*, vol. 6, no. 3, pp. 514–521, 2024, https://doi.org/10.47233/jteksis.v6i3.1310
[14] R. A. Husen, R. Astuti, L. Marlia, Rahmaddeni, and L. Efrizoni, "Analisis sentimen opini publik pada Twitter terhadap Bank BSI menggunakan algoritma machine learning," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 2, pp. 211–218, 2023, https://doi.org/10.57152/malcom.v3i2.901
[15] A. F. Pangestu, B. Rahmat, and A. N. Sihananto, "Analisis sentimen pada media sosial X terhadap implementasi Kurikulum Merdeka menggunakan metode FastText dan Long Short-Term Memory (LSTM)," *JIPI*, vol. 9, no. 4, pp. 2271–2280, 2024, https://doi.org/10.29100/jipi.v9i4.5665
[16] L. Damayanti and K. M. Lhaksmana, "Sentiment analysis of the 2024 Indonesia presidential election on Twitter," *Sinkron: Jurnal dan Penelitian Teknik Informatika*, vol. 8, no. 2, pp. 938–946, 2024, https://doi.org/10.33395/sinkron.v8i2.13379
[17] M. J. Islam, R. Datta, and A. Iqbal, "Actual rating calculation of the Zoom Cloud Meetings app using user reviews on Google Play Store with sentiment annotation of BERT and hybridization of RNN and LSTM," *Expert Systems with Applications*, vol. 223, Art. no. 119919, 2023, https://doi.org/10.1016/j.eswa.2023.119919
[18] A. Rehman, K. Das, Kamlish, and F. Abid, "Sentiment analysis using bidirectional encoder representations from transformers," *Proceedings of the Pakistan Academy of Sciences: A. Physical and Computational Sciences*, vol. 61, no. 2, pp. 153–165, 2024, https://doi.org/10.53560/PPASA(61-2)870
[19] K. Kusnawi and K. Anam, "Comparison of sentiment labeling using TextBlob, VADER, and Flair in public opinion analysis post-2024 presidential inauguration with IndoBERT," *Jurnal Teknik Informatika (JUTIF)*, vol. 6, no. 2, pp. 803–818, Apr. 2025, https://doi.org/10.52436/1.jutif.2025.6.2.4015
[20] S. L. Mirtaheri, A. Pugliese, N. Movahed, and R. Shahbazian, "A comparative analysis on using GPT and BERT for automated vulnerability scoring," *Intelligent Systems with Applications*, vol. 26, Art. no. 200515, 2025, https://doi.org/10.1016/j.iswa.2025.200515
[21] D. R. Alghifari, M. Edi, and L. Firmansyah, "Implementasi Bidirectional LSTM untuk analisis sentimen terhadap layanan Grab Indonesia," *Jurnal Manajemen Informatika (JAMIKA)*, vol. 12, no. 2, pp. 89–99, 2022, https://doi.org/10.34010/jamika.v12i2.7764
[22] S. Nachwa et al., "Pendekatan klasifikasi dalam knowledge discovery untuk analisis sentimen berbasis aspek pada ulasan Bandara Sultan Mahmud Badaruddin II di Google Maps," *JATI: Jurnal Mahasiswa Teknik Informatika*, vol. 9, no. 3, pp. 4782–4789, 2025, https://doi.org/10.36040/jati.v9i3.13776
[23] Y. Romadhoni and K. F. H. Holle, "Analisis sentimen terhadap PERMENDIKBUD No. 30 pada media sosial Twitter menggunakan metode Naive Bayes dan LSTM," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 7, no. 2, pp. 118–124, 2022, https://doi.org/10.30591/jpit.v7i2.3191
[24] L. Rofiqi and M. Akbar, "Analisis sentimen terkait RUU Perampasan Aset dengan Support Vector Machine," *JEKIN: Jurnal Teknik Informatika*, vol. 4, no. 3, pp. 529–538, 2024, https://doi.org/10.58794/jekin.v4i3.824
[25] S. P. Azzahra, Y. A. Apriyanto, and A. Wijaya, "Analisis sentimen ulasan aplikasi DeepL pada Google Play dengan metode Support Vector Machine (SVM)," *Jurnal Sistem Informasi (JUSIN)*, vol. 4, no. 2, pp. 59–66, 2023, https://doi.org/10.32546/jusin.v4i2.2368
[26] N. P. V. D. Saraswati, N. Yudistira, and P. P. Adikara, "Analisis sentimen terhadap perundungan siber pada Twitter menggunakan algoritma Bidirectional Encoder Representations from Transformer (BERT)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 2, pp. 909–916,2023.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 33
This work is licensed under a Creative Commons Attribution 4.0 International License.

ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 25-34
DOI: 10.47065/bulletinds.v6i1.11071
https://ejurnal.seminar-id.com/index.php/bulletinds

[27] A. Supian, B. T. Revaldo, N. Marhadi, L. Efrizoni, and Rahmaddeni, "Perbandingan kinerja Naive Bayes dan SVM pada analisis sentimen Twitter Ibukota Nusantara," *Jurnal Ilmiah Informatika*, vol. 12, no. 1, pp. 15–21, 2024, https://doi.org/10.33884/jif.v12i01.8721

[28] M. Yasir and R. Suraji, "Perbandingan metode klasifikasi Naive Bayes, Decision Tree, Random Forest terhadap analisis sentimen kenaikan biaya haji 2023 pada media sosial YouTube," *Jurnal Cahaya Mandalika*, vol. 3, no. 2, pp. 180–192, 2023, https://doi.org/10.36312/jcm.v3i2.1520

[29] J. E. Bata, Istiadi, and S. W. Irianandа, "Analisis sentimen ulasan pengunjung wisata Laguna Waikuri pada Google Maps menggunakan multi kelas Support Vector Machine," *JATI: Jurnal Mahasiswa Teknik Informatika*, vol. 8, no. 6, pp. 11115–11123, 2024, https://doi.org/10.36040/jati.v8i6.11229

[30] D. S. Utami, A. Erfina, and Mupaat, "Analisis sentimen ulasan terkait UNESCO Global Geopark di Google Maps dengan algoritma Naive Bayes," *J-SAKTI: Jurnal Sains Komputer dan Informatika*, vol. 6, no. 2, pp. 1154–1170, 2022.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 34
This work is licensed under a Creative Commons Attribution 4.0 International License.