 ISSN:2302-8149
e-ISSN:2540-9719

Llama dan *Gemma* vs *IndoBERT* dan *SBERT*: Analisis Sentimen *Zero-Shot* dan *Few-Shot*

Llama and Gemma vs IndoBERT and SBERT: Zero-Shot and Few-Shot Sentiment Analysis

[1]Haydar Yahya Asya'bani Nur*, [2]Muhammad Iqbal Akbar
[1,2]Program Studi Teknik Informatika, Fakultas Teknik, Universitas Negeri Malang
[1,2]Jl. Semarang No. 5, Kota Malang, Jawa Timur, Indonesia
*e-mail: haydar.yahya.2305356@students.um.ac.id

(***received***: 11 October 2026, ***revised***: 28 October 2026, ***accepted***: 5 November 2026)

Abstrak

Large Language Model (LLM) generatif seperti Llama dan Gemma menawarkan kemampuan zero-shot dan few-shot tanpa memerlukan fine-tuning, namun belum banyak penelitian yang membandingkannya secara langsung dengan model encoder berbasis embedding-similarity seperti IndoBERT dan SBERT dalam kondisi evaluasi yang setara, khususnya untuk sentiment analysis multibahasa Indonesia-Inggris. Penelitian ini bertujuan mengevaluasi dan membandingkan performa keempat model tersebut dalam skema zero-shot dan few-shot tanpa fine-tuning. Penelitian menggunakan dataset paralel NusaX-senti yang dibagi menjadi 80% train-pool dan 20% test-pool secara stratified, dengan enam contoh few-shot dan seratus skenario uji. Performa diukur menggunakan accuracy, precision, recall, dan F1-score makro, dan dilengkapi analisis confusion matrix. Perbedaan antarkondisi diuji dengan uji McNemar dan interval kepercayaan bootstrap. Hasil penelitian menunjukkan bahwa few-shot meningkatkan F1-score pada seluruh model, dengan peningkatan yang signifikan secara statistik pada empat dari tujuh kombinasi model dan bahasa. Peningkatan terbesar teramati pada IndoBERT (F1 0,297 menjadi 0,616), sedangkan peningkatan pada SBERT kecil (ΔF1 $\approx$ 0,07) dan tidak signifikan. F1-score tertinggi dicapai Llama-3.2-3B pada skema few-shot berbahasa Inggris (0,780; IK 95% 0,696-0,857), dengan catatan 23-24% respons few-shot Llama terpotong oleh batas 60 token meskipun rating tetap dapat diekstraksi dari seluruh respons. SBERT menunjukkan variasi F1 paling kecil antarkondisi (0,511-0,619), tetapi stabilitas ini disertai performa yang lebih rendah daripada Llama. Perbedaan performa antara Bahasa Indonesia dan Inggris tidak signifikan pada seluruh model. Temuan ini bermanfaat bagi pengembangan sentiment analysis Bahasa Indonesia yang efisien tanpa pelatihan model.

Kata kunci: few-shot learning, model bahasa generatif, multibahasa Indonesia-Inggris, embedding kemiripan semantik, analisis sentimen

Abstract

Generative Large Language Models (LLMs), such as Llama and Gemma, offer zero-shot and few-shot capabilities without requiring fine-tuning. However, few studies have directly compared these models with embedding-similarity-based encoder models, such as IndoBERT and SBERT, under equivalent evaluation conditions, particularly for Indonesian–English multilingual sentiment analysis. This study aims to evaluate and compare the performance of these four models under zero-shot and few-shot settings without fine-tuning. The study uses the parallel NusaX-senti dataset, which was stratified and divided into an 80% training pool and a 20% test pool, with six few-shot examples and 100 evaluation scenarios. Model performance was assessed using accuracy, precision, recall, and macro-F1 scores, supplemented by confusion matrix analysis. Differences across experimental conditions were examined using McNemar's test and bootstrap confidence intervals. The results show that few-shot prompting improved F1 scores across all models, with statistically significant improvements in four of the seven model–language combinations. The largest improvement was observed for IndoBERT, whose F1 score increased from 0.297 to 0.616, whereas SBERT showed a smaller improvement (ΔF1 $\approx$ 0.07) that was not statistically significant. The highest F1 score was achieved by Llama-3.2-3B under the English few-shot setting (0.780; 95% CI: 0.696–0.857). However, 23–24% of Llama's few-shot responses were truncated due to the 60-token limit, although sentiment ratings could still be

 ISSN:2302-8149
e-ISSN:2540-9719

extracted from all responses. SBERT exhibited the lowest variation in F1 scores across experimental conditions (0.511–0.619), although this stability was accompanied by lower performance than Llama. Performance differences between Indonesian and English were not statistically significant for any of the models. These findings provide insights into the development of computationally efficient Indonesian sentiment analysis systems without requiring model training.
Keywords: *few-shot learning, generative language model, Indonesian-English multilingual, semantic similarity embedding, sentiment analysis*

1. Pendahuluan

Perkembangan teknologi digital telah mendorong pertumbuhan data tekstual secara eksponensial di media sosial, situs ulasan produk, dan forum daring, sehingga bidang Natural Language Processing (NLP) mengembangkan sentiment analysis untuk mengklasifikasikan kecenderungan emosional suatu teks ke dalam kategori positif, negatif, atau netral. Penerapannya pada konteks Indonesia memiliki tantangan tersendiri, karena Bahasa Indonesia tergolong low-resource language dengan ketersediaan dataset dan model yang jauh lebih terbatas dibandingkan Bahasa Inggris [1]. Di sisi lain, Bahasa Inggris merupakan bahasa dengan sumber daya paling melimpah, sehingga membandingkan performa model yang sama pada teks Bahasa Indonesia dan padanannya dalam Bahasa Inggris penting untuk mengetahui sejauh mana kemampuan model dapat dialihkan ke bahasa dengan sumber daya terbatas. Dalam beberapa tahun terakhir, Large Language Model (LLM) generatif seperti Llama dan Gemma menawarkan paradigma baru berupa kemampuan zero-shot dan few-shot learning (in-context learning) [2] tanpa memerlukan fine-tuning, sementara model encoder seperti BERT dan IndoBERT [1] sejak sebelumnya telah dimanfaatkan melalui representasi embedding untuk klasifikasi berbasis cosine similarity, namun efektivitasnya bergantung pada kualitas embedding, karena embedding yang tidak dioptimalkan untuk kemiripan semantik dapat menghasilkan similarity yang kurang bermakna, sebuah fenomena yang dikenal sebagai anisotropy [3], [4], yang kemudian coba diatasi melalui Sentence-BERT (SBERT) dan contrastive learning [5]. IndoBERT tanpa fine-tuning sengaja disertakan sebagai baseline karena merupakan model encoder monolingual Bahasa Indonesia yang paling banyak digunakan pada penelitian analisis sentimen di Indonesia. Model ini mewakili embedding "mentah" yang belum dilatih untuk kemiripan kalimat, sehingga perbandingannya dengan SBERT, sesama model encoder Transformer yang dilatih secara kontrastif untuk kemiripan semantik, memungkinkan pengaruh tujuan pelatihan embedding diamati secara langsung. Pertanyaan yang menjadi perhatian penelitian ini adalah, di tengah pesatnya perkembangan LLM generatif, apakah pendekatan embedding klasik seperti IndoBERT dan SBERT masih kompetitif ketika dievaluasi dalam skema zero-shot dan few-shot yang setara, tanpa fine-tuning?

Berbagai penelitian terdahulu telah berupaya menjawab pertanyaan seputar kemampuan LLM dalam tugas sentiment analysis. Buscemi dan Proverbio [6] misalnya, membandingkan performa ChatGPT, Gemini, dan LLaMA dalam skema zero-shot pada sepuluh bahasa berbeda, namun penelitian tersebut belum menjangkau Bahasa Indonesia dan tidak menguji skema few-shot. Hasan et al. [7] membandingkan strategi zero-shot dan few-shot pada bahasa dengan sumber daya terbatas, yaitu Bahasa Bangla, dan memberikan kontribusi penting berupa temuan bahwa performa few-shot lebih baik dibandingkan zero-shot, namun penelitian tersebut membandingkan LLM dengan model yang di fine-tune, alih-alih dengan pendekatan embedding-similarity murni. Sementara itu, penelitian yang secara khusus mengevaluasi Llama maupun Gemma dalam skema zero-shot dan few-shot untuk sentiment analysis pada umumnya masih terbatas pada bahasa lain di luar Bahasa Indonesia, seperti Cielen et al. [8] yang mengevaluasi Llama dan Gemma bersama beberapa small language model lain pada teks berbahasa Inggris, Di Palma et al. [9] yang menyelidiki representasi sentimen pada model Llama melalui pendekatan probing, serta Cherakhloo et al. [10] yang membandingkan performa zero-shot dan few-shot berbagai LLM open-source, termasuk Gemma dan Llama, pada Bahasa Persia. Sepengetahuan penulis, belum ada penelitian yang menyandingkan Llama dan Gemma secara langsung dengan pendekatan embedding berbasis IndoBERT dan SBERT dalam kerangka eksperimen yang konsisten pada konteks Bahasa Indonesia. Kesenjangan (research gap) inilah yang mendasari urgensi dilakukannya penelitian ini.

Berdasarkan uraian di atas, permasalahan yang diangkat dalam penelitian ini adalah bagaimana perbandingan performa antara model generatif berbasis prompting (Llama dan Gemma) dengan model

 ISSN:2302-8149
e-ISSN:2540-9719

encoder berbasis embedding-similarity (IndoBERT dan SBERT) dalam menyelesaikan tugas sentiment analysis multibahasa Indonesia-Inggris, dengan skema zero-shot dan few-shot yang diterapkan secara konsisten pada kedua kelompok model tersebut, tanpa melibatkan proses fine-tuning pada model manapun. Evaluasi dilakukan menggunakan dataset NusaX-senti, sebuah dataset paralel Indonesia-Inggris yang telah divalidasi oleh penutur asli [11]. Permasalahan ini penting untuk dikaji secara mendalam mengingat kedua paradigma model tersebut memiliki karakteristik, kebutuhan sumber daya, serta mekanisme kerja yang berbeda, sehingga pemahaman komparatif mengenai keunggulan dan keterbatasan masing-masing pendekatan akan memberikan kontribusi baik secara akademik maupun praktis bagi pengembangan sistem sentiment analysis Bahasa Indonesia ke depannya.

Penelitian ini bertujuan untuk mengevaluasi dan membandingkan performa Llama dan Gemma sebagai representasi model generatif dengan IndoBERT dan SBERT sebagai representasi model encoder berbasis embedding, dalam menyelesaikan tugas sentiment analysis pada teks berbahasa Indonesia dan Inggris, dengan skema zero-shot dan few-shot prompting. Hasil penelitian ini diharapkan dapat memberikan kontribusi, baik secara teoretis maupun praktis. Secara teoretis, penelitian ini memperkaya khazanah literatur mengenai perbandingan paradigma model generatif dan embedding-similarity dalam konteks sentiment analysis multibahasa, khususnya untuk pasangan bahasa Indonesia-Inggris yang belum banyak dieksplorasi pada penelitian sebelumnya. Secara praktis, temuan penelitian ini dapat menjadi rujukan bagi pengembang maupun peneliti yang membutuhkan solusi sentiment analysis yang efisien dan siap pakai tanpa memerlukan sumber daya komputasi besar untuk pelatihan model, sehingga dapat mendukung pengambilan keputusan berbasis data secara lebih cepat dan terjangkau, misalnya dalam pemantauan kepuasan pelanggan, analisis reputasi merek, maupun riset opini publik berbasis media digital.

2. Tinjauan Literatur

Upaya mengevaluasi kemampuan Large Language Model (LLM) dalam tugas sentiment analysis multibahasa berkembang pesat, salah satunya oleh Buscemi dan Proverbio [6], yang membandingkan ChatGPT, Gemini, dan LLaMA pada sepuluh bahasa melalui skema zero-shot tervalidasi human-benchmark, namun belum mencakup Bahasa Indonesia maupun skema few-shot. Hasan et al. [7] melengkapi hal ini dengan membandingkan zero-shot dan few-shot pada Bahasa Bangla, ditemukan bahwa few-shot mengungguli zero-shot meski tetap di bawah model fine-tuned seperti BanglaBERT namun perbandingan tersebut menyandingkan LLM dengan model fine-tuned, bukan embedding-similarity tanpa pelatihan. Tren serupa turut diterapkan pada bahasa dengan sumber daya terbatas lainnya, seperti Bahasa Serbia [12] dan konteks tweet berbahasa Jerman [13], serta dikaji secara sistematis dalam survei teknik prompt engineering kontemporer [14]. Namun demikian, tidak semua penelitian mendukung superioritas few-shot secara mutlak beberapa studi menemukan bahwa penambahan contoh dapat menurunkan performa akibat over-prompting [15], dan bahkan skema zero-shot terbukti lebih akurat dibandingkan few-shot pada tugas klasifikasi berbasis LLM tertentu [16], sebuah nuansa yang perlu dipertimbangkan dalam interpretasi hasil penelitian ini.

Pada model berukuran kecil seperti Llama dan Gemma, Cielen et al. [8] menunjukkan potensinya sebagai alternatif LLM komersial dalam skema zero-shot, Di Palma et al. [9] mengungkap kesenjangan antara representasi internal dan kemampuan artikulasi sentimen Llama melalui teknik probing, dan Cherakhloo et al. [10] mendapati Gemma 2 unggul secara agregat di antara sebelas LLM open-source, termasuk Llama, pada berbagai tugas Bahasa Persia, meskipun pada tugas analisis sentimen Gemma tidak selalu menjadi model terbaik. Ketiganya tidak menyasar Bahasa Indonesia maupun menyandingkan Llama dan Gemma dengan pendekatan embedding. Pada jalur paradigma embedding-similarity, penelitian di Indonesia umumnya berfokus pada penerapan IndoBERT dengan fine-tuning penuh untuk domain spesifik, seperti analisis sentimen pemilu [17], kombinasi dengan model deep learning lain [18], perbandingan lintas bahasa daerah [19], maupun topik ekonomi terkini [20] sejalan dengan tren serupa pada bahasa lain yang juga mengandalkan fine-tuning BERT untuk sentiment analysis [21] seluruhnya menunjukkan performa kompetitif IndoBERT, namun selalu melalui pelatihan ulang, bukan skema zero-shot/few-shot berbasis embedding tanpa pelatihan seperti pada penelitian ini. Keterbatasan cosine similarity pada embedding yang tidak dioptimalkan untuk kemiripan semantik [3], [4] turut mendasari pengembangan SBERT dan contrastive learning [5],

 ISSN:2302-8149
e-ISSN:2540-9719

namun, sepengetahuan penulis, belum ada penelitian yang membandingkan langsung performa IndoBERT dan SBERT tanpa fine-tuning terhadap LLM generatif dalam skema zero-shot dan few-shot yang setara.

Berdasarkan sintesis di atas, dapat diidentifikasi bahwa penelitian yang secara bersamaan yaitu membandingkan model generatif berbasis prompting (Llama dan Gemma) dengan model encoder berbasis embedding-similarity (IndoBERT dan SBERT) tanpa fine-tuning pada kedua kelompok model, menerapkan skema zero-shot dan few-shot yang konsisten, dan berfokus pada konteks bahasa Indonesia-Inggris menggunakan dataset paralel tervalidasi [11], sepengetahuan penulis hingga saat ini belum dilakukan. Artikel ini secara khusus difokuskan untuk mengisi ketiga aspek kesenjangan tersebut secara simultan. Ringkasan posisi penelitian ini terhadap lima penelitian acuan utama disajikan pada Tabel 1.

Tabel 1 Ringkasan penelitian terdahulu dan kesenjangan penelitian (research gap)

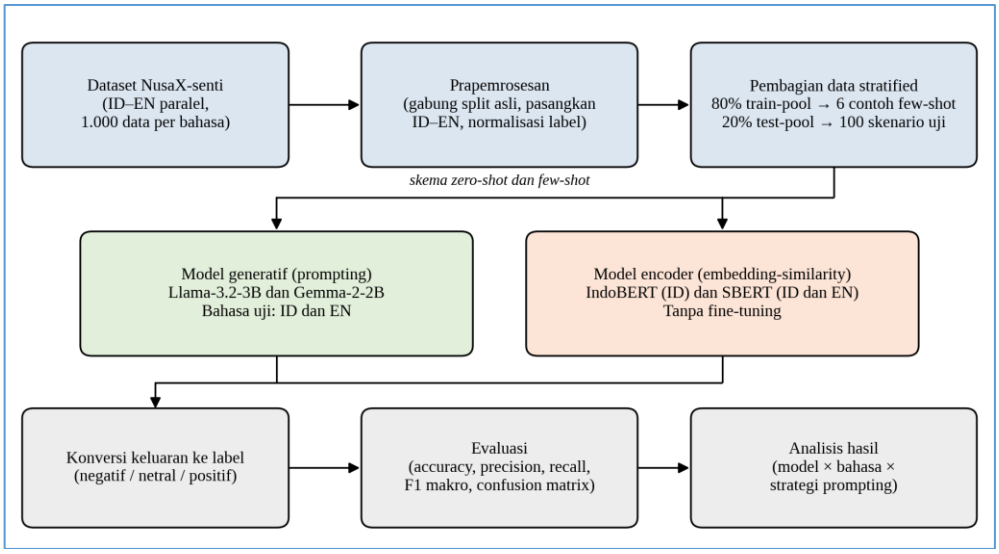
No	Peneliti & Tahun	Model/Metode	Temuan Utama	Gap Terhadap Penelitian Ini
1	Buscemi & Proverbio [6] (2024)	Zero-shot prompting, 4 LLM generatif, 10 bahasa, validasi human-benchmark	Performa LLM bervariasi antar bahasa; seluruh model kesulitan menangkap ironi	Murni zero-shot; tidak ada Bahasa Indonesia; tidak melibatkan Gemma maupun model encoder
2	Hasan et al. [7] (2024)	Zero-shot & few-shot LLM vs model fine-tuned, Bahasa Bangla	Few-shot mengungguli zero-shot; model fine-tuned tetap paling unggul	Dibandingkan dengan model fine-tuned, bukan embedding-similarity tanpa pelatihan; bukan Bahasa Indonesia
3	Cielen et al. [8] (2026)	Zero-shot, ansambel 5 SLM (termasuk Llama, Gemma), Bahasa Inggris	SLM berpotensi jadi alternatif LLM komersial yang lebih murah/privat	Hanya Bahasa Inggris; tidak ada few-shot; tidak ada model encoder embedding
4	Di Palma et al. [9] (2025)	Llama 1B/3B/8B, zero-shot/few-shot/probing	Classifier berbasis probing mengungguli Llama pada zero/few-shot langsung	Hanya Llama (tidak ada Gemma); hanya Bahasa Inggris; tidak ada IndoBERT/SBERT
5	Cherakhloo et al. [10] (2025)	Zero-shot & few-shot, 11 LLM open-source (termasuk Gemma 2 dan Llama 3.x), Bahasa Persia	Gemma 2 unggul secara agregat; pada tugas sentimen tidak selalu terbaik	Fokus Bahasa Persia; tidak melibatkan model encoder embedding

Berdasarkan kajian di atas, penelitian ini menguji tiga hipotesis. H1: strategi few-shot menghasilkan F1-score yang lebih tinggi daripada zero-shot pada setiap model, sebagaimana ditemukan Hasan et al. [7], meskipun beberapa studi melaporkan hasil sebaliknya [15], [16]. H2: kesenjangan performa antara Bahasa Indonesia dan Bahasa Inggris berbeda besaran dan arahnya antarmodel, mengingat Buscemi dan Proverbio [6] menemukan performa LLM yang bervariasi antarbahasa. H3: pada skema zero-shot, IndoBERT tanpa fine-tuning menghasilkan performa yang lebih rendah daripada SBERT karena embedding-nya tidak dioptimalkan untuk kemiripan semantik [3]-[5]. Ketiga hipotesis diuji pada Bab 4 menggunakan prosedur statistik yang dijelaskan pada Subbab 3.5.

 ISSN:2302-8149
e-ISSN:2540-9719

3. Metode Penelitian

3.1 Alur Penelitian



Gambar 1 Alur penelitian

Penelitian ini merupakan penelitian eksperimental kuantitatif yang membandingkan empat model pada tugas klasifikasi sentimen tiga kelas (positif, negatif, netral) untuk teks berbahasa Indonesia dan Inggris, yaitu Llama dan Gemma sebagai model generatif berbasis prompting serta IndoBERT dan SBERT sebagai model encoder berbasis embedding-similarity. Alur penelitian ditunjukkan pada Gambar 1. Dataset NusaX-senti dimuat dan diprapemroses, kemudian dibagi menjadi train-pool dan test-pool sebagai sumber contoh few-shot dan skenario uji. Keempat model dijalankan pada skenario uji dalam skema zero-shot dan few-shot, keluarannya dikonversi menjadi label sentimen, lalu dibandingkan dengan label acuan menggunakan metrik evaluasi dan dianalisis berdasarkan model, bahasa, dan strategi. Seluruh eksperimen dijalankan dengan bahasa pemrograman Python pada Google Colab berakselerator GPU, menggunakan pustaka Hugging Face Transformers dan Sentence-Transformers.

3.2 Dataset

Data yang digunakan adalah NusaX-senti [11], dataset analisis sentimen paralel yang dibangun dengan menerjemahkan dataset SmSA dari benchmark IndoNLU [22] ke bahasa lain melalui proses penerjemahan dan validasi oleh penutur asli. Penelitian ini memakai versi Bahasa Indonesia (sumber) dan Bahasa Inggris (terjemahan), masing-masing berjumlah 1.000 data yang merupakan gabungan split train (500), validation (100), dan test (400) pada rilis aslinya. Setiap data memiliki salah satu dari tiga label sentimen, yaitu negatif (383 data), positif (378 data), dan netral (239 data). Karena bersifat paralel, setiap kalimat berbahasa Indonesia memiliki pasangan berbahasa Inggris dengan label yang identik, sehingga perbedaan performa antar bahasa tidak disebabkan oleh perbedaan isi maupun label data. Distribusi label pada setiap tahap pengambilan data disajikan pada Tabel 2.

Tabel 2 Distribusi label pada dataset dan sampel penelitian

Label	Dataset	Train-pool (80%)	Test-pool (20%)	Contoh few-shot	Skenario uji
Negatif	383	306	77	2	33
Netral	239	191	48	2	34
Positif	378	303	75	2	33
Total	**1.000**	**800**	**200**	**6**	**100**

 ISSN:2302-8149
e-ISSN:2540-9719

3.3 Preprocessing Data

Tahap prapemrosesan sengaja dibuat minimal karena seluruh model memiliki tokenizer subword sendiri yang dirancang untuk menangani teks mentah. Langkah yang dilakukan meliputi penggabungan ketiga split asli menjadi satu himpunan, pemasangan versi Indonesia dan Inggris berdasarkan urutan baris yang identik disertai verifikasi bahwa jumlah barisnya sama, serta normalisasi penulisan label (huruf kecil dan tanpa spasi berlebih) menjadi tiga kelas. Tidak dilakukan case folding, penghapusan stopword, stemming, maupun pembersihan tanda baca, karena tahapan tersebut dapat menghilangkan sinyal sintaktis dan ekspresif yang dimanfaatkan model. Pada IndoBERT, teks dipotong hingga maksimum 128 token, sedangkan SBERT mengikuti batas panjang bawaan modelnya.

Data kemudian dibagi menjadi 80% train-pool (800 data) dan 20% test-pool (200 data) secara stratified per label dengan seed acak 42. Rasio ini mengikuti konvensi yang umum dipakai mengingat belum ada panduan tunggal mengenai rasio pembagian yang optimal [23]; porsi data uji 20% juga sama dengan yang digunakan Hasan et al. [7]. Perlu ditegaskan bahwa istilah train tidak merujuk pada pelatihan bobot model karena penelitian ini tidak melibatkan fine-tuning. Train-pool berfungsi sebagai sumber contoh few-shot sebagaimana pada in-context learning [2], sedangkan test-pool menjadi sumber skenario evaluasi. Validation set tidak digunakan karena tidak ada penyetelan hiperparameter, sejalan dengan [7] yang memakai set pengembangan hanya untuk model yang di-fine-tune.

Contoh few-shot berjumlah enam, diambil dari train-pool secara stratified (dua contoh per label), mengikuti kerangka N-way K-shot dengan N = 3 dan K = 2 sebagaimana pada prototypical networks [24]. Contoh yang sama digunakan pada seluruh model dan seluruh skenario uji, dengan bahasa yang disesuaikan dengan bahasa teks uji. Skenario uji berjumlah 100 data yang diambil dari test-pool dengan jumlah seimbang per kelas (33 data per kelas ditambah satu data acak, sehingga diperoleh 33 negatif, 34 netral, dan 33 positif), bukan proporsional terhadap distribusi aslinya. Pengambilan seimbang dipilih agar setiap kelas memiliki jumlah data yang cukup untuk menghitung recall per kelas dan menganalisis confusion matrix secara andal, karena dengan sampel proporsional kelas netral hanya terwakili sekitar 24 data. Konsekuensinya, nilai accuracy pada penelitian ini mencerminkan performa pada distribusi kelas yang seimbang, bukan pada distribusi alami dataset. Karena diambil dari test-pool, skenario uji tidak beririsan dengan contoh few-shot. Jumlah ini merupakan kompromi antara keterbatasan komputasi dan kecukupan sampel, dengan mengacu pada Pacchiardi et al. [25] yang menunjukkan bahwa himpunan rujukan sekitar 100 instance dapat dimanfaatkan untuk memprediksi performa LLM pada data yang belum dilihat. Pada ukuran ini, margin of error estimasi akurasi diperkirakan sekitar ±6,9% (tingkat kepercayaan 95%, asumsi proporsi 0,5, dengan koreksi populasi terbatas terhadap test-pool 200 data).

3.4 Model Penelitian

Penelitian ini menggunakan empat model pralatih tanpa fine-tuning (Tabel 3). Llama-3.2-3B-Instruct [26] dan Gemma-2-2B-it [27] dipilih sebagai model generatif berukuran kecil yang dapat dijalankan pada GPU terbatas. IndoBERT (indobenchmark/indobert-base-p1) [22] merupakan model encoder monolingual Bahasa Indonesia sehingga hanya dievaluasi pada teks berbahasa Indonesia, sedangkan SBERT [28] dipakai dalam versi multibahasa paraphrase-multilingual-MiniLM-L12-v2 [29] yang dapat mengevaluasi kedua bahasa. Perbedaan skala parameter antara kedua kelompok (miliaran berbanding ratusan juta) merupakan karakteristik paradigma yang dibandingkan dan diperlakukan sebagai batasan penelitian.

Tabel 3 Konfigurasi model dan skema zero-shot/few-shot

Model	ID Hugging Face	Paradigma	Bahasa uji	Zero-shot	Few-shot
Llama-3.2-3B	meta-llama/Llama-3.2-3B-Instruct	Generatif (prompting)	ID, EN	Instruksi tanpa contoh	Instruksi + 6 contoh
Gemma-2-2B	google/gemma-2-2b-it	Generatif (prompting)	ID, EN	Instruksi tanpa contoh	Instruksi + 6 contoh

 ISSN:2302-8149
e-ISSN:2540-9719

Model	ID Hugging Face	Paradigma	Bahasa uji	Zero-shot	Few-shot
IndoBERT	indobenchmark/indobert-base-p1	Encoder (embedding)	ID	Kemiripan ke deskripsi label	Kemiripan ke centroid 6 contoh
SBERT	paraphrase-multilingual-MiniLM-L12-v2	Encoder (embedding)	ID, EN	Kemiripan ke deskripsi label	Kemiripan ke centroid 6 contoh

Pada model generatif, tugas dirumuskan sebagai pemberian rating sentimen skala 1-10 (1 sangat negatif, 5-6 netral, 10 sangat positif) dengan format keluaran rating angka, diadaptasi dari rancangan prompt pada [6]. Pada skema zero-shot, prompt hanya berisi instruksi dan teks uji, sedangkan pada skema few-shot prompt disisipi enam contoh pasangan teks rating (negatif = 2, netral = 5, positif = 9). Pembangkitan teks menggunakan greedy decoding (do_sample = False) dengan peran sistem sebagai analis sentimen; pada Gemma, instruksi sistem digabungkan ke giliran pengguna karena template obrolannya tidak mendukung peran sistem. Panjang keluaran dibatasi maksimum 60 token baru, karena format jawaban yang diminta ("RATING: <angka>") hanya membutuhkan kurang dari sepuluh token; respons yang mencapai batas tersebut dicatat sebagai respons terpotong (truncated). Rating diekstraksi dari pola "RATING: <angka>" (atau "<angka>/10") dan dikonversi menjadi label dengan aturan rating ≤ 4 negatif, 5-6 netral, dan ≥ 7 positif. Respons yang tidak memuat pola rating tersebut dicatat sebagai gagal parse. Untuk memastikan determinisme, setiap skenario pada model generatif dijalankan tiga kali. Ketiga pengulangan menghasilkan prediksi yang identik, sehingga seluruh analisis menggunakan satu pengulangan dengan n = 100 skenario per kombinasi model, bahasa, dan strategi. Model encoder bersifat deterministik sehingga cukup dijalankan satu kali.

Pada model encoder, setiap teks diubah menjadi vektor embedding. SBERT menghasilkan embedding melalui fungsi encode pustaka Sentence-Transformers, sedangkan IndoBERT menghasilkannya melalui mean pooling terhadap hidden state lapisan terakhir dengan memperhitungkan attention mask. Seluruh vektor dinormalisasi terlebih dahulu, kemudian kemiripan antar vektor diukur menggunakan cosine similarity, yaitu metrik yang mengukur kedekatan arah antara dua vektor embedding tanpa dipengaruhi oleh panjang vektornya. Meskipun cosine similarity dapat kurang informatif pada embedding yang tidak dioptimalkan untuk kemiripan semantik [3], [4], metrik ini dipertahankan agar kedua model encoder dievaluasi dengan prosedur yang identik. Pada skema zero-shot, prediksi ditentukan dengan membandingkan embedding teks uji terhadap embedding kalimat deskripsi tiap label pada bahasa yang bersangkutan (misalnya “Teks ini menunjukkan sentimen yang positif dan menyenangkan.”), dan kelas dengan nilai cosine similarity tertinggi dipilih sebagai prediksi. Pada skema few-shot, setiap kelas diwakili oleh sebuah centroid, yaitu vektor rata-rata dari embedding seluruh contoh few-shot pada kelas tersebut, mengadaptasi prinsip prototypical networks [24], prediksi kemudian ditentukan oleh kelas yang centroid-nya paling mirip dengan embedding teks uji.

3.5 Metrik Evaluasi dan Uji Statistik

Label prediksi dibandingkan dengan label acuan NusaX-senti menggunakan empat metrik evaluasi standar untuk klasifikasi, yaitu accuracy, precision, recall, dan F1-score. Accuracy menunjukkan proporsi prediksi yang tepat dari seluruh skenario uji yang berhasil diproses; respons yang gagal parse dikeluarkan dari perhitungan metrik dan dilaporkan sebagai tingkat gagal parse. Precision menunjukkan proporsi prediksi suatu kelas yang benar-benar tepat, sedangkan recall menunjukkan proporsi data suatu kelas yang berhasil terdeteksi dengan benar oleh model. F1-score merupakan rata-rata harmonis dari precision dan recall. Ketiga metrik tersebut dihitung per kelas, lalu dirata-ratakan tanpa pembobotan (macro-average). Macro-average dipilih karena memberi bobot yang sama pada ketiga kelas, sehingga kegagalan model pada satu kelas tertentu, misalnya kelas netral yang paling sulit dikenali, tetap tercermin pada nilai akhir dan tidak tertutupi oleh performa yang baik pada kelas lain. Hal ini penting karena accuracy dapat tetap tampak wajar meskipun model hampir tidak pernah mengenali kelas netral dengan benar.

 ISSN:2302-8149
e-ISSN:2540-9719

Seluruh metrik dilaporkan untuk setiap kombinasi model, bahasa, dan strategi prompting, dilengkapi confusion matrix untuk menelaah pola kesalahan klasifikasi. Perbedaan antarkondisi diuji dengan uji McNemar eksak, yang sesuai untuk membandingkan dua pengklasifikasi pada data uji yang sama [30], menggunakan pasangan prediksi benar dan salah pada 100 skenario yang sama. Tiga kelompok perbandingan diuji, yaitu few-shot dan zero-shot pada setiap model dan bahasa (H1), Bahasa Indonesia dan Inggris pada setiap model dan strategi (H2), serta SBERT dan IndoBERT pada teks Bahasa Indonesia (H3). Nilai p dikoreksi dengan metode Holm [31] di dalam setiap kelompok untuk mengendalikan kesalahan akibat perbandingan berganda, dengan tingkat signifikansi $\alpha = 0,05$. Untuk menggambarkan ketidakpastian estimasi akibat ukuran sampel, interval kepercayaan (IK) 95% untuk F1-score dan untuk selisih F1-score antarkondisi diestimasi dengan metode bootstrap persentil [32] sebanyak 2.000 kali resampling skenario uji.

4. Hasil dan Pembahasan

4.1 Hasil Pembagian dan Pengambilan Sampel Data

Pembagian data menghasilkan train-pool 800 data dan test-pool 200 data dengan distribusi label yang proporsional, serta enam contoh few-shot dan 100 skenario uji sebagaimana dirinci pada Tabel 2. Contoh few-shot dan skenario uji yang sama digunakan pada seluruh model dan kedua bahasa, sehingga perbedaan performa yang teramati pada subbab berikutnya tidak disebabkan oleh perbedaan data uji. Meskipun demikian, perbedaan tersebut tetap dapat dipengaruhi oleh faktor lain di luar kemampuan model, seperti pemilihan contoh few-shot, rumusan prompt dan deskripsi label, serta skala parameter model (Subbab 4.5).

4.2 Hasil Evaluasi Performa Model

Tabel 4 Hasil evaluasi keempat model (n = 100 skenario per kombinasi)

Model	Bahasa	Strategi	n	Akurasi	Presisi	Recall	F1	IK 95% F1	Tingkat terpotong
Gemma-2-2B	EN	Few-shot	100	0,660	0,658	0,665	0,612	[0,517; 0,702]	0,00
Gemma-2-2B	EN	Zero-shot	100	0,480	0,436	0,484	0,430	[0,342; 0,518]	0,00
Gemma-2-2B	ID	Few-shot	100	0,730	0,776	0,734	0,703	[0,614; 0,784]	0,00
Gemma-2-2B	ID	Zero-shot	100	0,540	0,573	0,544	0,475	[0,381; 0,567]	0,00
IndoBERT	ID	Few-shot	100	0,620	0,634	0,622	0,616	[0,516; 0,712]	–
IndoBERT	ID	Zero-shot	100	0,330	0,281	0,332	0,297	[0,227; 0,364]	–
Llama-3.2-3B	EN	Few-shot	100	0,780	0,784	0,781	0,780	[0,696; 0,857]	0,23
Llama-3.2-3B	EN	Zero-shot	100	0,590	0,702	0,593	0,568	[0,463; 0,664]	0,00
Llama-3.2-3B	ID	Few-shot	100	0,730	0,742	0,733	0,718	[0,627; 0,802]	0,24
Llama-3.2-3B	ID	Zero-shot	100	0,600	0,728	0,604	0,563	[0,460; 0,658]	0,00
SBERT-Multilingual	EN	Few-shot	100	0,630	0,627	0,631	0,619	[0,518; 0,713]	–
SBERT-Multilingual	EN	Zero-shot	100	0,610	0,589	0,615	0,550	[0,468; 0,637]	–
SBERT-Multilingual	ID	Few-shot	100	0,590	0,591	0,592	0,578	[0,475; 0,674]	–
SBERT-Multilingual	ID	Zero-shot	100	0,580	0,574	0,585	0,511	[0,424; 0,589]	–

Tabel 4 menunjukkan bahwa F1-score skema few-shot lebih tinggi daripada zero-shot pada seluruh tujuh kombinasi model dan bahasa. Namun, uji McNemar (Tabel 5) menunjukkan bahwa peningkatan tersebut signifikan secara statistik setelah koreksi Holm hanya pada empat kombinasi, yaitu Gemma-2-2B pada kedua bahasa, Llama-3.2-3B pada Bahasa Inggris, dan IndoBERT. Pada Llama-3.2-3B berbahasa Indonesia, peningkatan F1 sebesar 0,155 (IK 95% 0,049-0,261) signifikan sebelum koreksi ($p = 0,019$) tetapi tidak setelah koreksi Holm ($p = 0,058$), sedangkan pada SBERT peningkatannya kecil ($\Delta F1 \approx 0,07$) dan tidak signifikan pada kedua bahasa. Dengan demikian, H1 hanya didukung sebagian. Peningkatan terbesar teramati pada IndoBERT, dengan F1-score naik dari 0,297 menjadi 0,616. Hal ini mengindikasikan bahwa beberapa contoh berlabel, baik sebagai konteks prompt maupun sebagai centroid embedding, membantu sebagian besar model mengarahkan prediksinya, tetapi manfaatnya kecil pada SBERT yang embedding-nya sudah dioptimalkan untuk kemiripan semantik.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 5 Uji McNemar few-shot terhadap zero-shot (n = 100 skenario berpasangan)

Model	Bahasa	F1 zero-shot	F1 few-shot	ΔF1	IK 95% ΔF1	Hanya few-shot benar	Hanya zero-shot benar	p (Holm)	Signifikan
Llama-3.2-3B	ID	0,563	0,718	+0,155	[0,049; 0,261]	20	7	0,058	Tidak
Llama-3.2-3B	EN	0,568	0,780	+0,212	[0,108; 0,324]	25	6	0,004	Ya
Gemma-2-2B	ID	0,475	0,703	+0,228	[0,129; 0,326]	21	2	< 0,001	Ya
Gemma-2-2B	EN	0,430	0,612	+0,182	[0,078; 0,281]	20	2	< 0,001	Ya
IndoBERT	ID	0,297	0,616	+0,319	[0,213; 0,425]	36	7	< 0,001	Ya
SBERT-Multilingual	ID	0,511	0,578	+0,067	[−0,041; 0,170]	14	13	1,000	Tidak
SBERT-Multilingual	EN	0,550	0,619	+0,069	[−0,050; 0,184]	19	17	1,000	Tidak

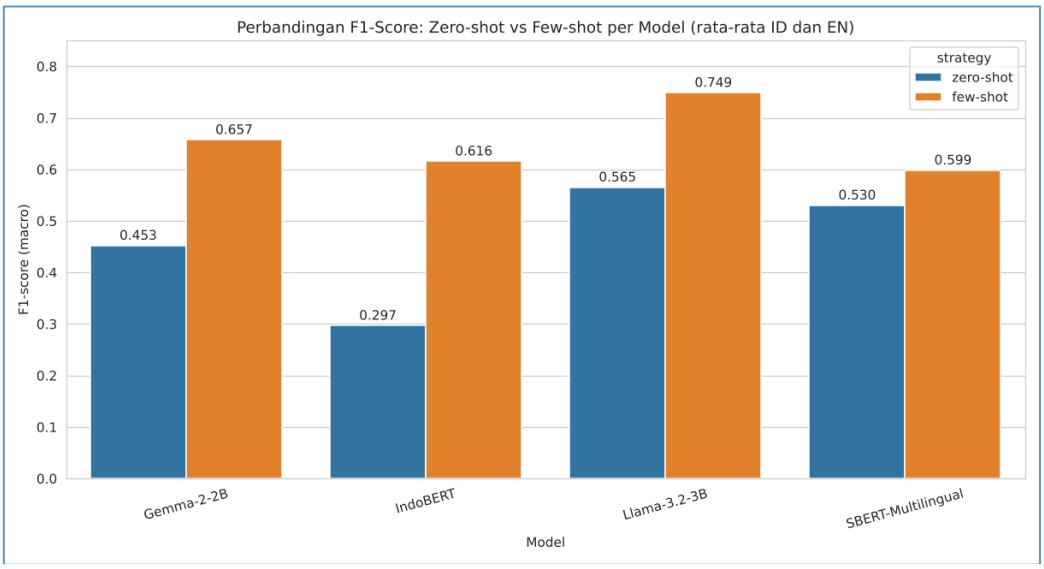
Performa IndoBERT pada skema zero-shot (accuracy 0,33, F1 0,297) patut mendapat perhatian khusus karena nilainya mendekati baseline tebakan acak untuk klasifikasi tiga kelas. Temuan ini sejalan dengan H3 mengenai keterbatasan cosine similarity pada embedding yang tidak dioptimalkan untuk kemiripan semantik [3], [4], karena IndoBERT tidak dilatih dengan tujuan tersebut, embeddingnya cenderung kurang mampu membedakan deskripsi antar label pada skema zero-shot. Hal ini diperkuat oleh perbandingan langsung dengan SBERT pada kondisi identik, yang mencapai F1 0,511 dan signifikan lebih tinggi daripada IndoBERT (uji McNemar, $p < 0,001$), meski keduanya sama-sama tidak di-fine-tune. Menariknya, pada skema few-shot keduanya tidak berbeda secara signifikan (F1 0,616 berbanding 0,578; $p = 0,69$), sehingga keunggulan SBERT hanya tampak ketika tidak ada contoh berlabel.

Kesenjangan performa antara Bahasa Indonesia dan Bahasa Inggris menunjukkan arah yang tidak seragam antarmodel. Gemma-2-2B memperoleh F1 lebih tinggi pada Bahasa Indonesia pada kedua skema, Llama-3.2-3B lebih tinggi pada Bahasa Inggris pada skema few-shot (F1 0,780 berbanding 0,718), sedangkan SBERT multibahasa sedikit lebih tinggi pada Bahasa Inggris dengan rentang F1 0,511-0,619 pada seluruh kondisinya. Namun, tidak satu pun perbedaan antarbahasa tersebut signifikan secara statistik (uji McNemar; seluruh p setelah koreksi Holm $\geq 0,71$ dan seluruh p tanpa koreksi $\geq 0,12$), sehingga H2 tidak didukung oleh data pada ukuran sampel 100 skenario dan arah kesenjangan di atas hanya dapat dibaca sebagai kecenderungan deskriptif. IndoBERT tidak disertakan dalam perbandingan ini karena hanya dievaluasi pada Bahasa Indonesia.

Secara keseluruhan, Llama-3.2-3B mencapai F1-score tertinggi, yaitu 0,780 (IK 95% 0,696-0,857) pada skema few-shot berbahasa Inggris. Keunggulan ini tidak berlaku untuk seluruh model generatif: pada skema few-shot berbahasa Inggris, Gemma-2-2B (F1 0,612) setara dengan SBERT (F1 0,619), sedangkan pada Bahasa Indonesia Gemma-2-2B (F1 0,703) berada di bawah Llama namun di atas kedua model encoder. Interval kepercayaan antarmodel juga sebagian tumpang-tindih dan perbedaan antarmodel tidak diuji secara formal, sehingga keunggulan Llama perlu dibaca sebagai kecenderungan pada sampel ini. Llama-3.2-3B juga satu-satunya model dengan respons terpotong (23-24% pada skema few-shot). Tanda "–" pada kolom tingkat terpotong di Tabel 4 menunjukkan bahwa metrik ini tidak berlaku bagi IndoBERT dan SBERT, karena model encoder tidak menghasilkan teks. Karena rating tetap dapat diekstraksi dari seluruh respons tersebut (tingkat gagal parse 0%), keterpotongan ini tidak mengurangi jumlah data yang dievaluasi.

 ISSN:2302-8149
e-ISSN:2540-9719

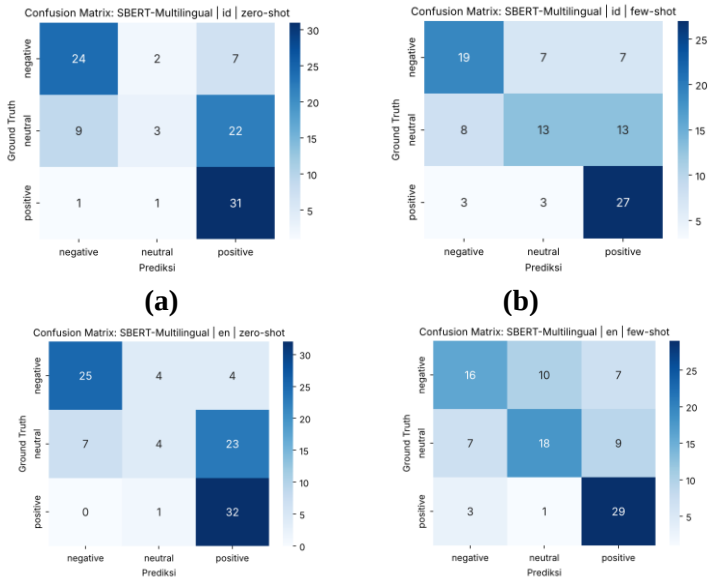
4.3 Visualisasi Perbandingan Zero-Shot dan Few-Shot



Gambar 2 Perbandingan F1-score zero-shot dan few-shot per model

Gambar 2 menyajikan perbandingan visual F1-score antara skema zero-shot dan few-shot pada keempat model, dirata-ratakan lintas kedua bahasa. Grafik ini mempertegas pola yang telah diidentifikasi sebelumnya: batang few-shot lebih tinggi dibandingkan batang zero-shot pada seluruh model, sejalan dengan Tabel 4 yang menunjukkan peningkatan F1 pada seluruh tujuh kombinasi model dan bahasa, meskipun tidak semua peningkatan tersebut signifikan secara statistik. Besaran selisih keduanya bervariasi, dengan IndoBERT menunjukkan kesenjangan paling lebar antara kedua skema (F1 zero-shot mendekati 0,30, melonjak ke atas 0,60 pada few-shot), sementara SBERT-Multilingual menunjukkan kesenjangan paling sempit di antara keempat model (selisih sekitar 0,07), mengindikasikan performanya relatif lebih stabil terhadap perubahan strategi prompting dibandingkan ketiga model lainnya. Secara deskriptif, Llama-3.2-3B memiliki rata-rata F1-score tertinggi pada kedua skema.

4.4 Analisis Kesalahan Klasifikasi melalui Confusion Matrix



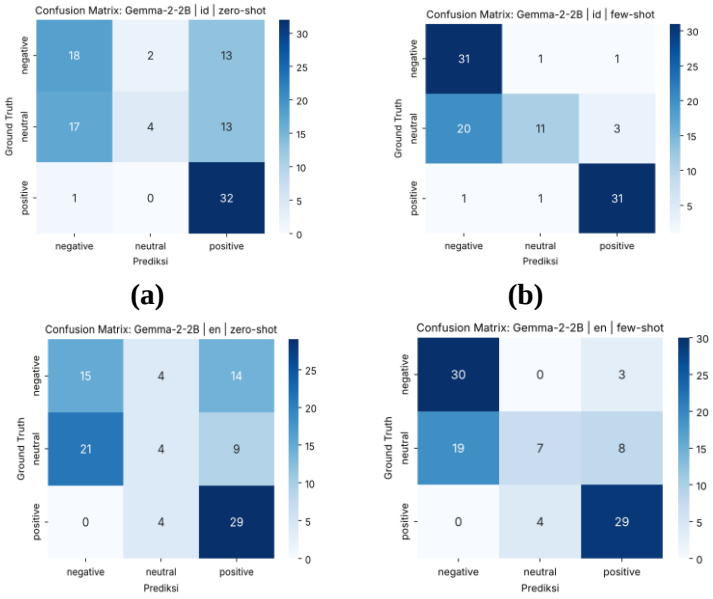
(a) **(b)**

(c) **(d)**

Gambar 3 Confusion Matrix SBERT-Multilingual: (a) ID, zero-shot; (b) ID, few-shot; (c) EN, zero-shot; (d) EN, few-shot

 ISSN:2302-8149
e-ISSN:2540-9719

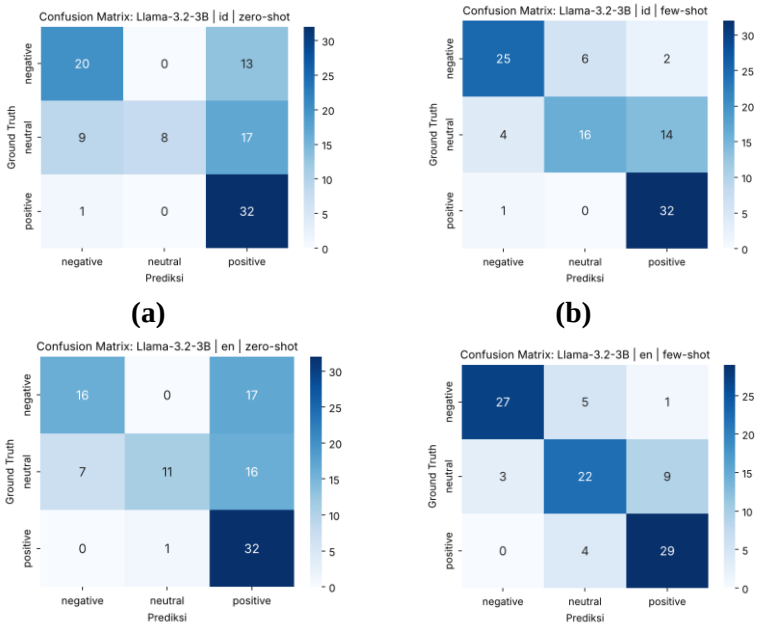
Gambar 3 menyajikan confusion matrix SBERT-Multilingual pada keempat kombinasi bahasa dan strategi, yang secara konsisten menunjukkan kelas netral sebagai yang paling sulit dikenali dan kelas positif sebagai yang paling mudah dikenali. Pada skema zero-shot, model menunjukkan bias prediksi yang kuat ke arah kelas positif recall netral hanya 9% pada Bahasa Indonesia (3 dari 34 data) dan 12% pada Bahasa Inggris (4 dari 34 data) dengan sebagian besar data netral justru terklasifikasi sebagai positif (22 dari 34 data pada Bahasa Indonesia dan 23 dari 34 data pada Bahasa Inggris). Skema few-shot memperbaiki bias ini secara nyata, meningkatkan recall netral menjadi 38% pada Bahasa Indonesia dan 53% pada Bahasa Inggris. Menariknya, perbaikan yang cukup besar pada kelas netral ini tidak terlalu terlihat pada angka accuracy keseluruhan (hanya naik tipis dari 0,58 menjadi 0,59 pada Bahasa Indonesia, dan dari 0,61 menjadi 0,63 pada Bahasa Inggris), karena accuracy mengukur ketepatan secara agregat sehingga perbaikan pada kelas minoritas seperti netral kurang tercermin dibanding bila dilihat per kelas. Pola ini menunjukkan bahwa pemberian contoh few-shot secara spesifik mengurangi kecenderungan model mengklasifikasikan teks ambigu sebagai sentimen positif, sebuah bias yang muncul pada SBERT ketika tidak diberi contoh pembanding.



Gambar 4 Confusion Matrix Gemma-2-2B: (a) ID, zero-shot; (b) ID, few-shot; (c) EN, zero-shot; (d) EN, few-shot

Gambar 4 menyajikan confusion matrix Gemma-2-2B pada keempat kombinasi. Berbeda dari SBERT yang bias ke kelas positif, Gemma pada skema zero-shot justru kesulitan mengenali kelas netral karena prediksinya terpecah hampir merata ke kelas negatif dan positif (recall netral hanya 11,8% pada kedua bahasa), sementara kelas positif sudah dikenali sangat baik bahkan tanpa contoh (recall 88-97%). Skema few-shot memperbaiki terutama kelas negatif (recall naik dari sekitar 45-55% menjadi 91-94%) dan netral (naik menjadi 20,6% pada Bahasa Inggris dan 32,4% pada Bahasa Indonesia), sejalan dengan kenaikan accuracy keseluruhan dari 0,48 menjadi 0,66 (Bahasa Inggris) dan 0,54 menjadi 0,73 (Bahasa Indonesia). Pola ini menunjukkan bahwa kelemahan utama Gemma pada zero-shot bukan bias ke satu kelas tertentu seperti SBERT, melainkan ketidakmampuan membedakan sentimen netral dari sentimen ekstrem, yang berkurang setelah model diberi contoh berlabel.

 ISSN:2302-8149
e-ISSN:2540-9719

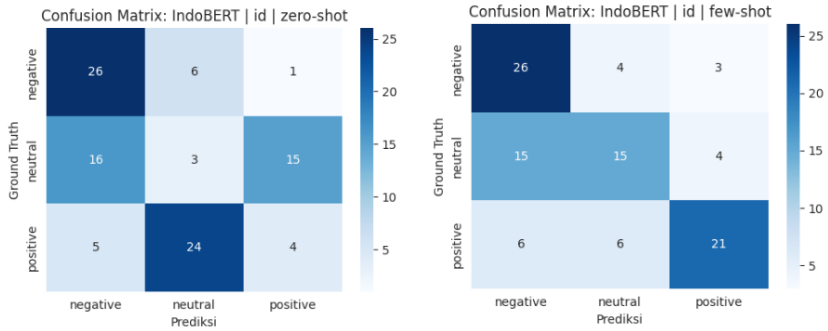


(a) **(b)**

(c) **(d)**

Gambar 5 Confusion Matrix Llama-3.2-3B: (a) ID, zero-shot; (b) ID, few-shot; (c) EN, zero-shot; (d) EN, few-shot

Gambar 5 menyajikan confusion matrix Llama-3.2-3B. Pada skema zero-shot, kelas netral juga paling sulit dikenali (recall 23,5% pada Bahasa Indonesia dan 32,4% pada Bahasa Inggris), dengan sekitar separuh data netral justru terklasifikasi positif, sementara kelas positif dikenali hampir sempurna (recall 97,0% pada kedua bahasa) bahkan tanpa contoh. Skema few-shot terutama memperbaiki kelas netral (recall naik menjadi 47,1% pada Bahasa Indonesia dan 64,7% pada Bahasa Inggris) dan kelas negatif (dari 60,6% menjadi 75,8% pada Bahasa Indonesia dan dari 48,5% menjadi 81,8% pada Bahasa Inggris), sementara recall positif tetap 97,0% pada Bahasa Indonesia dan turun menjadi 87,9% pada Bahasa Inggris. Perubahan ini sejalan dengan kenaikan accuracy dari 0,60 menjadi 0,73 (Bahasa Indonesia) dan dari 0,59 menjadi 0,78 (Bahasa Inggris), accuracy tertinggi di antara seluruh model dan kombinasi pada penelitian ini. Pola ini menunjukkan bahwa Llama sudah mengenali sentimen positif dengan baik tanpa contoh, dan few-shot berperan besar dalam meningkatkan kemampuannya mengenali sentimen netral dan negatif.



(a) **(b)**

Gambar 6 Confusion Matrix IndoBERT: (a) ID, zero-shot; (b) ID, few-shot

Gambar 6 menyajikan confusion matrix IndoBERT, satu-satunya model yang hanya dievaluasi pada Bahasa Indonesia. Pada skema zero-shot, pola kesalahannya berbeda dari ketiga model lain, IndoBERT justru paling banyak salah mengklasifikasikan data positif sebagai netral (24 dari 33 data, recall positif hanya 12,1%), sementara recall netral juga sangat rendah (8,8%); hanya kelas negatif yang dikenali cukup baik (recall 78,8%). Skema few-shot memperbaiki kedua kelas yang lemah tersebut secara drastis recall positif melonjak dari 12,1% menjadi 63,6% dan recall netral naik dari 8,8% menjadi 44,1% sementara recall negatif tetap stabil di 78,8%, sejalan dengan kenaikan accuracy keseluruhan dari 0,33 menjadi 0,62, peningkatan terbesar di antara seluruh model pada penelitian ini.

 ISSN:2302-8149
e-ISSN:2540-9719

Pola kebingungan utama IndoBERT pada zero-shot, yakni antara kelas positif dan netral (bukan antara sentimen berlawanan), mengindikasikan bahwa embedding yang dihasilkannya tanpa contoh pembanding kesulitan memisahkan sentimen yang searah namun berbeda intensitas, dan centroid few-shot efektif mengatasi kelemahan spesifik ini.

4.5 Kelebihan dan Keterbatasan Penelitian

Dibandingkan dengan kelima penelitian acuan pada Tabel 1, kelebihan penelitian ini terletak pada perbandingan langsung dua paradigma yang sebelumnya umumnya dikaji secara terpisah, yaitu model generatif berbasis prompting (Llama dan Gemma) dan model encoder berbasis embedding-similarity (IndoBERT dan SBERT), dalam skema zero-shot dan few-shot yang konsisten tanpa fine-tuning pada kedua kelompok model. Buscemi dan Proverbio [6] maupun studi evaluasi Llama dan Gemma lainnya [8], [10] hanya mengevaluasi model generatif, sedangkan penelitian berbasis IndoBERT di Indonesia [17]-[20] melibatkan fine-tuning penuh. Evaluasi tanpa pelatihan tambahan pada kedua paradigma memungkinkan keduanya dibandingkan dengan masukan, contoh, dan data uji yang sama.

Hasan et al. [7] melaporkan bahwa few-shot mengungguli zero-shot pada GPT-4 dan BLOOMZ untuk Bahasa Bangla. Penelitian ini menemukan arah peningkatan yang serupa pada model yang jauh lebih kecil (Llama-3.2-3B dan Gemma-2-2B, masing-masing sekitar 3 dan 2 miliar parameter), dengan peningkatan yang signifikan pada tiga dari empat kombinasi model generatif dan bahasa. Temuan ini mengindikasikan bahwa model generatif berukuran kecil yang dapat dijalankan pada GPU gratis berpotensi digunakan untuk sentiment analysis Bahasa Indonesia dan Inggris tanpa sumber daya komputasi sebesar model komersial.

Terakhir, penggunaan dataset paralel NusaX-senti [11] memungkinkan kesenjangan performa antarbahasa dianalisis pada kondisi terkontrol, karena setiap kalimat Bahasa Indonesia memiliki padanan Bahasa Inggris dengan label identik. Pada penelitian ini, arah kesenjangan bahasa tampak berbeda antarmodel: Gemma-2-2B cenderung lebih baik pada Bahasa Indonesia, sedangkan Llama-3.2-3B dan SBERT cenderung lebih baik pada Bahasa Inggris pada skema few-shot. Namun, tidak ada perbedaan antarbahasa yang signifikan secara statistik, sehingga pengujian dengan jumlah skenario yang lebih besar diperlukan untuk memastikan apakah kesenjangan tersebut nyata, sebagaimana variasi performa antarbahasa yang dilaporkan Buscemi dan Proverbio [6] pada LLM komersial.

Penelitian ini juga memiliki beberapa keterbatasan yang perlu dipertimbangkan dalam menginterpretasikan hasilnya. Pertama, evaluasi hanya menggunakan 100 skenario uji dengan margin of error sekitar ±6,9% dan satu rangkaian enam contoh few-shot yang diambil sekali dengan seed tetap, sehingga performa few-shot yang dilaporkan dapat berubah apabila contoh yang dipilih berbeda. Kedua, hasil bersumber dari satu dataset NusaX-senti dengan versi Bahasa Inggris berupa terjemahan, bukan teks asli penutur Bahasa Inggris, sehingga kesimpulan antar bahasa belum tentu berlaku pada data yang ditulis langsung dalam Bahasa Inggris. Ketiga, model yang dibandingkan berbeda skala parameter (miliaran pada Llama dan Gemma, ratusan juta pada IndoBERT dan SBERT), IndoBERT bersifat monolingual sehingga hanya dievaluasi pada Bahasa Indonesia, dan seluruh model berukuran kecil sehingga temuan belum tentu berlaku pada model yang lebih besar. Keempat, batas 60 token baru menyebabkan 23-24% respons Llama-3.2-3B pada skema few-shot terpotong. Meskipun rating tetap dapat diekstraksi dari seluruh respons (tingkat gagal parse 0%), keterpotongan ini menandakan bahwa model tidak selalu mematuhi format keluaran yang diminta, dan aturan ekstraksi rating yang berbeda dapat menghasilkan label yang berbeda. Kelima, dengan 100 skenario uji, uji statistik hanya mampu mendeteksi perbedaan yang cukup besar, sehingga selisih kecil seperti perbedaan antarbahasa atau peningkatan few-shot pada SBERT baru dapat dipastikan dengan sampel yang lebih besar.

5. Kesimpulan

Penelitian ini membandingkan Llama-3.2-3B dan Gemma-2-2B sebagai model generatif berbasis prompting dengan IndoBERT dan SBERT sebagai model encoder berbasis embedding-similarity untuk sentiment analysis teks Bahasa Indonesia dan Inggris, dalam skema zero-shot dan few-shot tanpa fine-tuning. Few-shot meningkatkan F1-score pada seluruh kombinasi model dan bahasa, dan

 ISSN:2302-8149
e-ISSN:2540-9719

peningkatan tersebut signifikan secara statistik (uji McNemar dengan koreksi Holm) pada Gemma-2-2B di kedua bahasa, Llama-3.2-3B pada Bahasa Inggris, dan IndoBERT. Peningkatan terbesar terjadi pada IndoBERT (F1 0,297 menjadi 0,616), sedangkan pada SBERT peningkatannya kecil dan tidak signifikan.

Llama-3.2-3B mencapai F1-score tertinggi (0,780 pada few-shot berbahasa Inggris), yang menunjukkan bahwa model generatif berukuran kecil dapat kompetitif untuk sentiment analysis tanpa pelatihan tambahan. Keunggulan ini tidak berlaku untuk seluruh model generatif: Gemma-2-2B unggul atas kedua model encoder pada Bahasa Indonesia, tetapi setara dengan SBERT pada Bahasa Inggris. Pada skema zero-shot, SBERT signifikan lebih baik daripada IndoBERT yang performanya mendekati tebakan acak, tetapi keduanya setara setelah diberi contoh few-shot. Perbedaan performa antara Bahasa Indonesia dan Bahasa Inggris tidak signifikan pada seluruh model. Penelitian lanjutan disarankan menggunakan jumlah skenario uji yang lebih besar, variasi contoh few-shot, serta model dan bahasa yang lebih beragam.

Referensi

[1] F. Koto, N. Aisyah, H. Li, and T. Baldwin, "*Large Language Models only Pass Primary School Exams in Indonesia: A Comprehensive Test on IndoMMLU*," in Proc. 2023 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2023, pp. 12359–12374, DOI: 10.18653/v1/2023.emnlp-main.760.

[2] Q. Dong et al., "*A Survey on in-Context Learning*," in Proc. 2024 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2024, pp. 1107–1128, DOI: 10.18653/v1/2024.emnlp-main.64.

[3] H. Steck, C. Ekanadham, and N. Kallus, "*Is Cosine-Similarity of Embeddings Really about Similarity?*," in Companion Proc. ACM Web Conference 2024, 2024, DOI: 10.1145/3589335.3651526.

[4] K. Zhou, K. Ethayarajh, D. Card, and D. Jurafsky, "*Problems with Cosine as a Measure of Embedding Similarity for High Frequency Words*," in Proc. 60th Annual Meeting of the Association for Computational Linguistics (ACL), Vol. 2, 2022, pp. 401–423, DOI: 10.18653/v1/2022.acl-short.45.

[5] T. Gao, X. Yao, and D. Chen, "*SimCSE: Simple Contrastive Learning of Sentence Embeddings*," in Proc. 2021 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2021, pp. 6894–6910, DOI: 10.18653/v1/2021.emnlp-main.552.

[6] A. Buscemi and D. Proverbio, "*ChatGPT vs Gemini vs LLaMA on Multilingual Sentiment Analysis*," arXiv preprint arXiv:2402.01715, 2024, DOI: 10.48550/arXiv.2402.01715.

[7] M. A. Hasan et al., "*Zero- and Few-Shot Prompting with LLMs: A Comparative Study with Fine-Tuned Models for Bangla Sentiment Analysis*," in Proc. 2024 Joint Int. Conf. Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italy, 2024, pp. 17808–17818.

[8] D. Cielen, K. W. De Bock, and L. Florès, "*Exploring Zero-Shot SLM Ensembles as an Alternative to LLMs for Sentiment Analysis*," Information Fusion, Vol. 126, pt. B, art. No. 103666, 2026, DOI: 10.1016/j.inffus.2025.103666.

[9] D. Di Palma, A. De Bellis, G. Servedio, V. W. Anelli, F. Narducci, and T. Di Noia, "*LLaMAs have Feelings Too: Unveiling Sentiment and Emotion Representations in LLaMA Models through Probing*," in Proc. 63rd Annual Meeting of the Association for Computational Linguistics (ACL), Vol. 1, 2025, pp. 6124–6142, DOI: 10.18653/v1/2025.acl-long.306.

[10] M. Cherakhloo, A. Abbasi, M. S. Sarafraz, and B. V. Vahdat, "*Benchmarking Open-Source Large Language Models for Persian in Zero-Shot and Few-Shot Learning*," arXiv preprint arXiv:2510.12807, 2025, DOI: 10.48550/arXiv.2510.12807.

[11] G. I. Winata et al., "*NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages*," in Proc. 17th Conf. European Chapter of the Association for Computational Linguistics (EACL), 2023, pp. 815–834, DOI: 10.18653/v1/2023.eacl-main.57.

[12] M. Ikonić Nešić, S. Petalinkar, M. Škorić, R. Stanković, and B. Rujević, "*Advancing Sentiment Analysis in Serbian Literature: A Zero and Few-Shot Learning Approach using the Mistral Model*," in Proc. 6th Int. Conf. Computational Linguistics in Bulgaria (CLIB 2024), 2024, pp. 58–70.

 ISSN:2302-8149
e-ISSN:2540-9719

[13] S. Münker, K. Kugler, and A. Rettinger, "*Zero-Shot Prompt-based Classification: Topic Labeling in Times of Foundation Models in German Tweets*," in Proc. 63rd Annual Meeting of the Association for Computational Linguistics (ACL), Student Research Workshop, 2025, pp. 53–63, DOI: 10.18653/v1/2025.acl-srw.4.
[14] S. Schulhoff et al., "*The Prompt Report: A Systematic Survey of Prompt Engineering Techniques*," arXiv preprint arXiv:2406.06608, 2024, DOI: 10.48550/arXiv.2406.06608.
[15] Y. Tang, D. Tuncel, C. Koerner, and T. Runkler, "*The Few-Shot Dilemma: Over-Prompting Large Language Models*," in Proc. 3rd Int. Conf. Foundation and Large Language Models (FLLM), 2025, pp. 134–141, DOI: 10.1109/FLLM67465.2025.11391015.
[16] B. P. Allen, F. Polat, and P. Groth, "*SHROOM-INDElab at SemEval-2024 Task 6: Zero- and Few-Shot LLM-based Classification for Hallucination Detection*," in Proc. 18th Int. Workshop on Semantic Evaluation (SemEval-2024), 2024, pp. 839–844, DOI: 10.18653/v1/2024.semeval-1.120.
[17] L. Geni, E. Yulianti, and D. I. Sensuse, "*Sentiment Analysis of Tweets before the 2024 Elections in Indonesia using BERT Language Models*," Jurnal Ilmiah Teknik Elektro Komputer dan Informatika, Vol. 9, No. 3, pp. 746–757, 2023, DOI: 10.26555/jiteki.v9i3.26490.
[18] D. G. Mandhasiya, H. Murfi, and A. Bustamam, "*The Hybrid of BERT and Deep Learning Models for Indonesian Sentiment Analysis*," Indonesian Journal of Electrical Engineering and Computer Science, Vol. 33, No. 1, pp. 591–602, 2024, DOI: 10.11591/ijeecs.v33.i1.pp591-602.
[19] T. D. Purnomo and J. Sutopo, "*Comparison of Pre-Trained BERT-based Transformer Models for Regional Language Text Sentiment Analysis in Indonesia*," International Journal Science and Technology, Vol. 3, No. 3, pp. 11–21, 2024, DOI: 10.56127/ijst.v3i3.1739.
[20] F. N. Khaqqi, L. Alfat, and I. Nurhaida, "*Sentiment Analysis of US-China Tariffs using IndoBERT and Economic Impact on Indonesia*," Journal of Applied Informatics and Computing, Vol. 9, No. 6, pp. 3000–3011, 2025, DOI: 10.30871/jaic.v9i6.11544.
[21] A. Bello, S.-C. Ng, and M.-F. Leung, "*A BERT Framework to Sentiment Analysis of Tweets*," Sensors, Vol. 23, No. 1, art. No. 506, 2023, DOI: 10.3390/s23010506.
[22] B. Wilie et al., "*IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding*," in Proc. 1st Conf. Asia-Pacific Chapter of the Association for Computational Linguistics and 10th Int. Joint Conf. on Natural Language Processing (AACL-IJCNLP), 2020, pp. 843–857, DOI: 10.18653/v1/2020.aacl-main.85.
[23] V. R. Joseph, "*Optimal Ratio for Data Splitting*," Statistical Analysis and Data Mining: The ASA Data Science Journal, Vol. 15, No. 4, pp. 531–538, 2022, DOI: 10.1002/sam.11583.
[24] J. Snell, K. Swersky, and R. Zemel, "*Prototypical Networks for Few-Shot Learning*," in Advances in Neural Information Processing Systems (NeurIPS), Vol. 30, 2017.
[25] L. Pacchiardi, L. G. Cheke, and J. Hernández-Orallo, "*100 Instances is All You Need: Predicting the Success of a New LLM on Unseen Data by Testing on a Few Instances*," arXiv preprint arXiv:2409.03563, 2024, DOI: 10.48550/arXiv.2409.03563.
[26] A. Grattafiori et al., "*The Llama 3 Herd of Models*," arXiv preprint arXiv:2407.21783, 2024, doi: 10.48550/arXiv.2407.21783.
[27] Gemma Team, "*Gemma 2: Improving Open Language Models at a Practical Size*," arXiv preprint arXiv:2408.00118, 2024, DOI: 10.48550/arXiv.2408.00118.
[28] N. Reimers and I. Gurevych, "*Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*," in Proc. 2019 Conf. Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3982–3992, DOI: 10.18653/v1/D19-1410.
[29] N. Reimers and I. Gurevych, "*Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation*," in Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 4512–4525, DOI: 10.18653/v1/2020.emnlp-main.365.
[30] T. G. Dietterich, "*Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms*," Neural Computation, Vol. 10, No. 7, pp. 1895–1923, 1998, DOI: 10.1162/089976698300017197.
[31] S. Holm, "*A Simple Sequentially Rejective Multiple Test Procedure*," Scandinavian Journal of Statistics, Vol. 6, No. 2, pp. 65–70, 1979.

 ISSN:2302-8149
e-ISSN:2540-9719

[32] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall, 1993.