Jurnal Teknologi Komputer dan Informasi (JUTEKINF)
Vol. 14, No. 1 Januari - Juni 2026, Hal. 10-24

Comparing Prompting Strategies for Low-Resource Indonesian Languages: GPT-4o-mini and IndoBERT on Javanese and Sundanese

Sandi Fadilah*[1], Ika Safitri Windiarti[2], , Gellysa Urva[3]
[1]Faculty of Business Management and Information Technology, Universiti Muhammadiyah Malaysia, Perlis, Malaysia. [2]Faculty of Business and Informatics, Universitas Muhammadiyah Palangkaraya, Palangka Raya, Indonesia. [3]Institut Teknologi dan Bisnis Riau Pesisir
e-mail: p5250078@student.umam.edu.my[1]*, ika.windiarti@umpr.ac.id[2], gellysa.urva@gmail.com[3]

Abstract

Javanese and Sundanese are two of Indonesia's most widely spoken regional languages, yet both suffer from severe underrepresentation in NLP research due to limited annotated data and few pre-trained resources. This study compares GPT-4o-mini and IndoBERT-base on three-class sentiment analysis for both languages using NusaX-Senti, evaluating zero-shot and few-shot prompting (nine examples) against an off-the-shelf fine-tuned classifier. Experiments used 100 stratified samples per language. GPT-4o-mini substantially outperforms IndoBERT-base: few-shot achieves Macro-F1 of 0.868 (Javanese) and 0.882 (Sundanese) versus 0.295 and 0.317 for IndoBERT ($p<0.001$, McNemar test). Few-shot prompting significantly improves Javanese performance ($\Delta F1=+0.122$, $p=0.015$), driven by the Neutral class (F1: 0.560→0.800), but shows no significant gain for Sundanese where zero-shot baseline already reaches 0.858. For Indonesian regional languages where domain-specific fine-tuning data is unavailable, prompted large language models offer a practical alternative to pre-trained BERT-based classifiers, requiring only minimal labeled examples rather than full retraining infrastructure.

Keywords: Sentiment Analysis, Low-Resource Languages, Few-Shot Prompting, Indonesian Regional Languages, Language Models

1. INTRODUCTION

Indonesia has more than 700 regional languages, many spoken by tens of millions of people, but poorly represented in NLP research (Koto et al., 2020). Javanese (~84 million speakers) and Sundanese (~34 million) are among the most widely spoken. Both are low-resource in NLP terms: annotated training data is scarce, dedicated pre-trained models are rare, and evaluation benchmarks are limited (Brown et al., 2020a; Koto et al., 2020; Winata et al., 2023). As user-generated content in regional languages grows on social media, demand for automated text analysis tools is increasing. Sentiment analysis classifying text by expressed polarity is among the most directly useful, with applications in opinion monitoring, product reviews, and social discourse analysis (Devlin et al., 2019; Nasution et al., 2025).

Two approaches currently dominate NLP for sentiment analysis. The first is the fine-tuned transformer, with BERT (Wu et al., 2025) and its Indonesian adaptation IndoBERT (Wali et al., 2026) as the main examples. IndoBERT, developed by Koto et al. on large Indonesian corpora, has become the standard baseline for Indonesian NLP and has been applied to sentiment taskszacross student feedback (Nasution et al., 2025), social media (Přibáň et al., 2024), and dynamic opinion analysis (Lossio-Ventura et al., 2024). The second is the generative large language model (LLM), represented most prominently by OpenAI's GPT series (Nazi et al., 2025a). Unlike fine-tuned models, LLMs can classify with prompting alone, via zero-shot or few-shot in-context learning, without labeled data or domain-specific training (OpenAI et al., 2023a). Fine-tuned models are more specialized but require labeled data and compute; LLMs are more flexible but sacrifice task-specific optimization (Sinaga et al., 2025; Zouidine & Khalil, 2025).

Comparative studies between LLMs and fine-tuned BERT models exist for Arabic (Alahmadi, 2025; Rusnachenko et al., 2024), Spanish (Nazi et al., 2025a), Russian (Kastrati et al., 2025), and English (Devlin et al., 2019; Zouidine & Khalil, 2025), and generally show that LLMs perform competitively or better in low-data settings (Kocoń et al., 2023; Suhartono et al.,

E-ISSN : 2830 – 7879 || P-ISSN : 2355 - 1887

2024). Indonesian work has mostly focused on Bahasa Indonesia (Jazuli et al., 2024; Winata et al., 2023), with little attention to regional languages. To our knowledge, limited prior work has directly compared GPT-4o-mini and IndoBERT on Javanese and Sundanese sentiment analysis using a standardized zero-shot and few-shot protocol on a shared benchmark. If a lightweight LLM performs well through prompting alone, it is immediately usable for regional languages where labeled data is hard to obtain.

This study compares GPT-4o-mini and IndoBERT-base on three-class sentiment analysis for Javanese and Sundanese, using NusaX-Senti (Koto et al., 2020) as the benchmark. NusaX-Senti covers ten Indonesian regional languages and received an Outstanding Paper award at EACL 2023. GPT-4o-mini is tested with zero-shot prompting (no examples) and few-shot prompting (nine labeled examples, three per class). IndoBERT-base is used as a fine-tuned baseline without additional domain-specific training. Statistical significance is assessed using the McNemar test on per-sample prediction arrays.

Three research questions guide the study: RQ1: Does GPT-4o-mini outperform IndoBERT-base without fine-tuning on Javanese and Sundanese sentiment analysis, as measured by Macro-F1? RQ2: Does few-shot prompting significantly improve GPT-4o-mini performance over zero-shot on these languages? RQ3: Are GPT-4o-mini's performance patterns consistent across Javanese and Sundanese?

The study makes four contributions: one of the first empirical comparisons of GPT-4o-mini and IndoBERT-base on these two languages using a shared public benchmark; per-class analysis showing the Neutral class as the primary challenge for both models; evidence that few-shot prompting gains are language-dependent, modulated by the model's zero-shot baseline; and reproducible results based on a publicly available dataset and public APIs. Section II describes the methodology; Section III presents the results and discussion; and Section IV concludes.

2. METODE

This study compares GPT-4o-mini and IndoBERT-base on sentiment analysis for two low-resource Indonesian regional languages, Javanese and Sundanese. GPT-4o-mini is evaluated under zero-shot and few-shot prompting; IndoBERT-base is used as a fine-tuned baseline. Figure 1 shows the overall research pipeline.

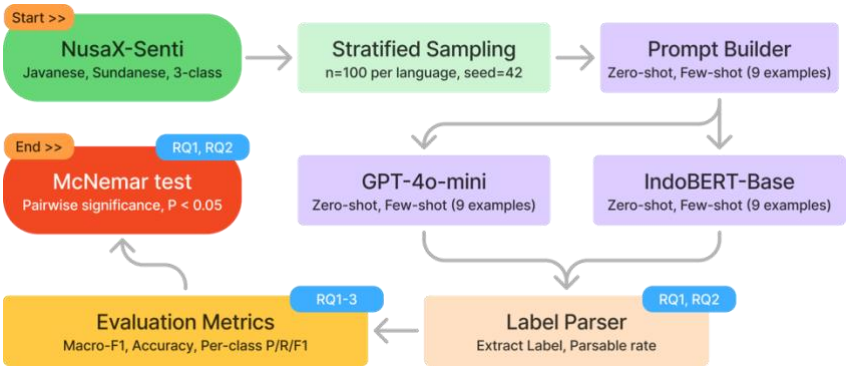

Figure 1. Overall Research Pipeline

A. Dataset

The study uses NusaX-Senti (Koto et al., 2020), a multilingual sentiment benchmark covering ten Indonesian regional languages, developed by Winata et al. and published as an Outstanding Paper at EACL 2023. The dataset is publicly available on HuggingFace under a CC-BY-SA 4.0 license with predefined train and test splits. NusaX-Senti was selected because it is the only publicly available, manually annotated sentiment dataset for Indonesian regional languages at scale. Its three-class label scheme (positive, neutral, negative) allows evaluation across polarity levels, and its predefined splits prevent data leakage during few-shot example selection.

Javanese (jav) and Sundanese (sun) were chosen as the two evaluation languages. Both are low-resource in NLP terms: limited annotated corpora exist, and no dedicated large-scale pre-

E-ISSN : 2830 – 7879 || P-ISSN : 2355 - 1887

trained models have been developed for either language (Koto et al., 2020). The pair also allows a cross-language comparison of whether performance patterns hold across two related but distinct regional languages. From each language's test split, 100 samples were drawn by stratified random sampling (seed=42), yielding 33-34 samples per sentiment class. This near-uniform class distribution matters because Macro-F1, the primary metric, weights each class equally regardless of support. Few-shot examples were drawn exclusively from the training split to prevent any overlap with the test set.

Table 1. Dataset configuration

Language	Code	n (test)	Positive	Neutral	Negative
Javanese	Jav	100	34	33	33
Sundanese	Sun	100	34	33	33
Total		200	68	66	66

Sample Size and Statistical Power

This study employs n=100 per language as an exploratory investigation of prompting strategies for low-resource sentiment analysis. This sample size provides:

1. Adequate power for primary comparison (RQ1): The GPT-4o-mini vs. IndoBERT-base comparison yielded 71 and 64 discordant pairs-well above the minimum threshold of 20-yielding highly significant results ($p<0.001$). Conclusions for RQ1 are supported by adequate statistical power.
2. Sufficient class representation: Stratified sampling ensures 33-34 instances per sentiment class, enabling reliable per-class F1 computation for the three-way classification task.
3. Consistency with pilot studies: Sample sizes of 100-150 per condition are common in exploratory NLP research for low-resource languages (Nasution et al., 2025; Nazi et al., 2025b; Winata et al., 2023).
4. Practical constraints: Larger samples would provide greater precision but were constrained by annotation verification time and API costs. Accordingly, all conclusions are framed as exploratory. RQ1 findings (large effect, disc ≥ 64) are sufficiently powered; RQ2 findings (disc=17 for Javanese, disc=8 for Sundanese) are treated as directional only and not generalized beyond this sample.

We acknowledge that n=100 per language limits statistical power for detecting small prompting effects (RQ2). Discordant pair counts of 17 (Javanese) and 8 (Sundanese) fall below the recommended threshold of 20, and both results should be interpreted as directional rather than conclusive. The statistically significant result for Javanese ($p=0.015$) is suggestive but not definitive. Future work with $n\geq200$ per language would enable reliable detection of smaller effect sizes (power ≥0.80).

B. Models

Two models were evaluated: a generative large language model and a fine-tuned encoder-based classifier. GPT-4o-mini is a lightweight, closed-source large language model developed by OpenAI (OpenAI et al., 2023b). It was accessed via the OpenAI commercial API using the model string gpt-4o-mini (sometimes displayed as GPT-4o-mini in interfaces, but the underlying model throughout all experiments is gpt-4o-mini), with temperature=0.0 and seed=42 for deterministic outputs. It was selected because it offers strong multilingual capabilities at substantially lower cost than larger GPT-4-class models. No fine-tuning was performed on NusaX-Senti; the model was evaluated purely through prompting.

IndoBERT-base is a BERT-based model pre-trained on Indonesian corpora by Koto et al. and fine-tuned for three-class sentiment classification. The model used is mdhugol/indonesia-bert-sentiment-classification, which is built on an IndoBERT-base backbone and fine-tuned for sentiment classification, and accessed via the HuggingFace Inference API. It represents the standard approach for Indonesian NLP tasks (Kocoń et al., 2023; Sinaga et al., 2025) and is used here as the fine-tuned baseline. This model was fine-tuned on a general Indonesian sentiment

E-ISSN : 2830 – 7879 || P-ISSN : 2355 - 1887

dataset, not on NusaX-Senti, so the comparison is between zero-resource prompting on one side and general-domain fine-tuning on the other.

Baseline Selection and Deployment Scenario

IndoBERT-base (mdhugol/indonesia-bert-sentiment-classification) was selected to represent a practical deployment scenario where no domain-specific retraining is feasible due to lack of labeled data or computational resources-a common constraint for Indonesian regional languages. This experimental design intentionally creates an asymmetric comparison: GPT-4o-mini uses zero-resource prompting (no training data required), while IndoBERT-base is deployed as a general-domain classifier (trained on Indonesian, not regional languages).

This reflects real-world conditions where practitioners must choose between: (a) API-based LLM prompting with minimal setup, or (b) deploying pre-trained BERT models without language-specific adaptation.

Important Clarification: The observed performance gap does not represent the architectural ceiling of BERT-based models. Prior work demonstrates that IndoBERT fine-tuned on domain-matched data achieves Macro-F1 > 0.80 on Indonesian sentiment tasks (Jazuli et al., 2024; Sinaga et al., 2025). Our IndoBERT baseline's low performance (F1: 0.295-0.317) stems from language mismatch (Bahasa Indonesia ≠ Javanese/Sundanese) rather than architectural limitations. We position this work as evaluating deployment readiness rather than architectural capacity. A fair architectural comparison would require fine-tuning IndoBERT on NusaX-Senti training data, or including multilingual baselines such as mBERT or XLM-R, which we propose as critical future work to establish the performance ceiling of encoder-based models for these languages.

Table 2. Model configurations

Model	Type	Access	Temperature	Seed
GPT-4o-mini	Closed-source LLM	OpenAI API	0.0	42
IndoBERT-Base	Fine-tuned BERT	HuggingFace Inference	N/A	N/A

C. Prompting Strategies

Two prompting strategies were applied to GPT-4o-mini, differing in the amount of task guidance the model receives at inference time. Zero-shot prompting gives the model a task instruction and the target text, with no labeled examples (OpenAI et al., 2023a). The system prompt specifies the classification task, the three sentiment labels, the target language, and a strict single-word output format, with no task-specific demonstrations. Few-shot prompting extends the zero-shot prompt with nine labeled examples (three per sentiment class) drawn from the NusaX-Senti training split using stratified random sampling with seed=42 (Nababan & Alan, 2023; OpenAI et al., 2023a). The examples appear in the user prompt before the target text to test whether in-context learning improves classification in low-resource languages.

Both strategies were also formally applied to IndoBERT-Base, but the model produced identical predictions regardless of prompt format. The HuggingFace Inference API classifier pipeline processes only the target text and ignores system instructions and few-shot examples, which is consistent with how encoder-based classifiers are typically deployed (Zouidine & Khalil, 2025). IndoBERT-Base, therefore, functions as a single fixed-condition baseline, and its results are reported identically under both strategies.

Table 3. Prompt structure for each strategy

Component	Zero-shot	Few-shot
System prompt	You are a sentiment classifier. Classify the sentiment of the following Javanese text as positive (positif), neutral (netral), or negative (negative). Respond with one word only.	

E-ISSN : 2830 – 7879 || P-ISSN : 2355 - 1887

Few-shot block	(none)	*apik tenan filme*, (positif), *biasa wae, ora ana sing spesial* (netral), *ora seneng karo layananane* (negative), [6 more examples, 2 per class, same format]
User prompt	Text: "[target text]" Sentiment:	Text: "[target text]" Sentiment:

D. Evaluation Metrics

The primary evaluation Metric is Macro-F1, the unweighted mean of F1-scores across all three sentiment classes. Each class is weighted equally regardless of frequency, which makes Macro-F1 appropriate for balanced three-class evaluation and consistent with related work in multilingual and low-resource sentiment analysis (Aji et al., 2022; Kocoń et al., 2023). Macro-F1 is defined as:

$$\text{Macro-F1} = \frac{1}{C}\sum_{c=1}^{C} F1_c \quad [1]$$

where C=3 is the number of sentiment classes and $F1_c$ is the F1-score for class c, computed as the harmonic mean of precision (P_c) and recall (R_c):

$$F1_c = 2 \times \frac{P_c \times R_c}{P_c + R_c} \quad [2]$$

Where:

$$P_c = \frac{TP_c}{TP_c + FP_c} \quad [3]$$

$$R_c = \frac{TP_c}{TP_c + FN_c} \quad [4]$$

With TP_c, FP_c, and FN_c denoting true positives, false positives, and false negatives for class c, respectively. Secondary metrics reported include overall accuracy, weighted F1-score, and per-class precision and recall, providing a detailed view of model behavior across individual sentiment categories. Statistical significance was assessed using the McNemar test (Kocoń et al., 2023), a non-parametric paired test that evaluates whether two classifiers make significantly different prediction errors on the same test instances. The McNemar statistic with Yates' continuity correction is:

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c} \quad [5]$$

where b is the count of samples correctly classified by model A but not model B, and c is the reverse. At least 20 discordant pairs (b + c >= 20) are required for reliable results (Kocoń et al., 2023), The significance threshold was $p < 0.05$.

E. Experimental Setup

All experiments were implemented in Python. GPT-4o-mini was accessed via the OpenAI SDK; IndoBERT-base via the HuggingFace Inference API. Raw outputs and per-sample predictions were logged to JSON files for each run. A label parser extracted the sentiment label from GPT-4o-mini's free-form text output, covering minor variation in capitalization, spacing,

E-ISSN : 2830 – 7879 || P-ISSN : 2355 - 1887

and punctuation. Parseable rate (the proportion of outputs that yielded a valid label) was recorded alongside F1 as a secondary quality check.

The full protocol comprised 400 GPT-4o-mini API calls (100 samples × 2 languages × 2 strategies), and 400 IndoBERT-base inference calls under equivalent conditions. Seed=42 was fixed for all sampling; temperature=0.0 for GPT-4o-mini; stratified sampling was used for both test set construction and few-shot selection. GPT-4o-mini achieved a parseable rate of 0.99–1.00 across all conditions, with at most one output per condition requiring fallback handling. IndoBERT-base produced structured outputs in all 400 inference calls (parseable rate: 1.00).

Reproducibility Protocol

All experiments used deterministic sampling (temperature=0.0, seed=42) to ensure exact reproducibility. This configuration: (1) eliminates stochastic variance by producing argmax tokens rather than sampled outputs, avoiding artifacts that could obscure prompting effects; (2) enables replication by allowing future researchers to reproduce our exact results using identical prompts and API configuration; (3) provides conservative estimates by returning the model's highest-confidence prediction; and (4) matches baseline conditions, as IndoBERT inference is deterministic.

We acknowledge that alternative prompt templates or example orderings may yield different results. However, our protocol isolates the effect of prompting strategy (zero-shot vs. few-shot) from generation stochasticity. Prior work on in-context learning (Brown et al., 2020b; Dong et al., 2024) shows that few-shot classification is relatively stable to moderate prompt variations when class labels are well-defined. Nonetheless, the use of a single prompt template and fixed example ordering limits the generalizability of our findings. Systematic prompt robustness analysis-including template ablation and varied example orderings-remains an important direction for future work.

3. RESULTS AND DISCUSSIONS

Table 4. Overall Results: Macro-F1, Weighted-F1, Accuracy, And Parseable Rate

Model	Strategy	Language	Macro-F1	Weighted-F1	Accuracy	Parseable
GPT-4o-mini	Zero-shot	Javanese	0.746	0.755	0.768	0.99
GPT-4o-mini	Few-shot	Javanese	0.868	0.872	0.870	1.00
GPT-4o-mini	Zero-shot	Sundanese	0.858	0.862	0.860	1.00
GPT-4o-mini	Few-shot	Sundanese	0.882	0.884	0.880	1.00
IndoBERT-Base	Zero-shot	Javanese	0.295	0.283	0.240	1.00
IndoBERT-Base	Few-shot	Javanese	0.295	0.283	0.240	1.00
IndoBERT-Base	Zero-shot	Sundanese	0.317	0.301	0.280	1.00
IndoBERT-Base	Few-shot	Sundanese	0.317	0.301	0.280	1.00

A. Overall Performance Comparison (RQ1)

Table 4 shows the full results across all models, strategies, and language conditions. GPT-4o-mini outperforms IndoBERT-base everywhere, by a large margin. Under few-shot prompting, GPT-4o-mini reaches Macro-F1 of 0.868 (Javanese) and 0.882 (Sundanese), gaps of 0.573 and 0.565 over IndoBERT-Base. Even zero-shot, it scores 0.746 and 0.858, still far above IndoBERT-Base's 0.295 and 0.317. Both models produced parseable outputs in 99-100% of cases throughout.

IndoBERT-Base's low scores need context. This comparison does not reflect the maximum achievable performance of IndoBERT, as the model was not fine-tuned on NusaX-Senti. Instead, it represents a practical deployment scenario where retraining is not feasible - a condition common in low-resource regional language settings. The model used here, mdhugol/indonesia-bert-sentiment-classification, was fine-tuned on a general Indonesian sentiment dataset rather

E-ISSN : 2830 – 7879 || P-ISSN : 2355 - 1887

than NusaX-Senti. This was intentional: the study tests IndoBERT as a ready-to-use classifier with no retraining, closer to an actual deployment scenario. These scores do not reflect the ceiling of the IndoBERT architecture. Prior work fine-tuning IndoBERT on domain-matched data reports Macro-F1 above 0.80 (Devlin et al., 2019; Lossio-Ventura et al., 2024); the gap here stems from a mismatch in the training distribution, not the model itself.

IndoBERT-base also produced identical predictions under zero-shot and few-shot conditions (disc=0). This is expected: the HuggingFace Inference API pipeline extracts only the target text and ignores prompt format (Zouidine & Khalil, 2025). RQ1 answer: GPT-4o-mini with no domain-specific fine-tuning substantially outperforms off-the-shelf IndoBERT-base on low-resource Javanese and Sundanese sentiment analysis.

B. Effect of Prompting Strategy (RQ2)

Figure 2 shows Macro-F1 across all models, strategies, and languages. For GPT-4o-mini, the benefit of few-shot prompting (nine labeled examples, three per class) differs substantially between the two languages.

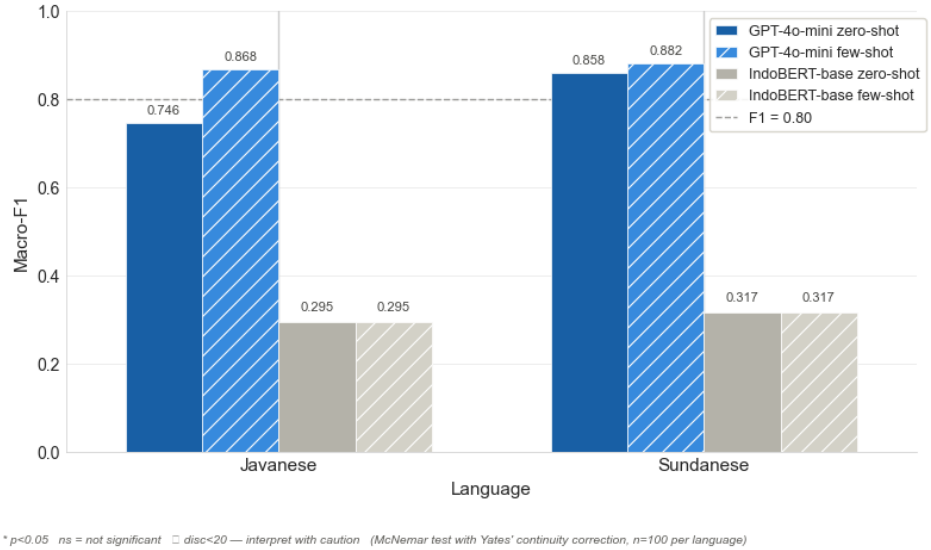

Figure 2. Macro-F1 scores by model, strategy, and language

On Javanese, few-shot prompting is helpful: Macro-F1 rises from 0.746 to 0.868, a gain of 0.122. The McNemar test (b=14, c=3, disc=17, p=0.015) yields p<0.05, though 17 discordant pairs fall below the recommended minimum of 20. This provides suggestive evidence of a prompting benefit but should be treated as a preliminary indication rather than a definitive significance claim.

On Sundanese, the gain is 0.024 (0.858 to 0.882), and the McNemar test (b=5, c=3, disc=8, p=0.724) is not significant. With only eight discordant pairs, the test has limited power here regardless. The difference between languages makes sense. GPT-4o-mini's zero-shot baseline on Sundanese (0.858) is already high, leaving little room for improvement, a ceiling effect observed in other multilingual prompting work (Dong et al., 2024; Nazi et al., 2025a). In Javanese, the baseline is lower (0.746), so in-context examples contribute more. This aligns with earlier findings that few-shot examples are most helpful when the model lacks strong priors for the target distribution (OpenAI et al., 2023a; Suhartono et al., 2024). For IndoBERT-Base, the prompting strategy did not affect either language (disc=0 in both cases). RQ2 answer: Few-shot prompting improves GPT-4o-mini on Javanese at a statistically significant level. It shows no significant benefit for Sundanese, where zero-shot performance is already near the ceiling.

C. Cross-Language Analysis (RQ3)

E-ISSN : 2830 – 7879 || P-ISSN : 2355 - 1887

GPT-4o-mini performs better on Sundanese than on Javanese, and the difference is most visible under zero-shot prompting. Sundanese zero-shot (0.858) leads Javanese (0.746) by 0.112 Macro-F1 points. With few-shot prompting, the gap narrows substantially: Javanese rises to 0.868 and Sundanese to 0.882, converging within 0.014 points. Figure 3 summarizes the average Macro-F1 across both prompting strategies, confirming that Sundanese (0.870) consistently outperforms Javanese (0.807) for GPT-4o-mini at the aggregate level, while IndoBERT-base remains uniformly low across both languages (0.295 and 0.317)

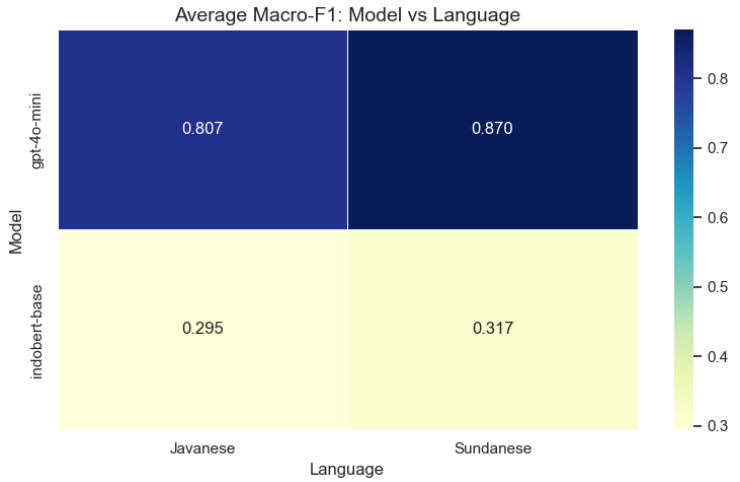

Figure 3. Average Macro-F1 by Model and Language

Figure 3. Average Macro-F1 by model and language (mean of zero-shot and few-shot, n=100 per language). GPT-4o-mini consistently outperforms IndoBERT-Base, with Sundanese yielding higher average scores than Javanese across both models.

Per-class results in Table 5 and Figure 4 pinpoint the source. Figure 4 visualizes the per-class F1 improvement from zero-shot to few-shot prompting for GPT-4o-mini across both languages. In zero-shot, Javanese Neutral F1 is 0.560, the lowest point across all conditions, compared to Sundanese Neutral at 0.781. When few-shot examples are introduced, Javanese Neutral rises to 0.800 (Δ=+0.240), largely closing the gap with Sundanese. Positive and Negative classes score comparably across both languages and both strategies, confirming that the cross-language difference is isolated to the Neutral class. The Negative class consistently achieves the highest F1 across all GPT-4o-mini conditions (0.912–0.943 under few-shot), suggesting negative sentiment carries the clearest lexical signal in both Javanese and Sundanese (Brown et al., 2020b; Nazi et al., 2025b).

Several linguistic factors may contribute to the higher Neutral difficulty in Javanese. Javanese employs an elaborate speech-level system (ngoko, madya, krama) that encodes social register within lexical choice, potentially creating ambiguity for sentiment classification when polarity cues are subtle or absent (Aji et al., 2022). Neutral utterances-which by definition lack strong polarity signals-may be particularly susceptible to misclassification when register variation alters surface form without changing sentiment. Additionally, Javanese text in social media frequently exhibits code-mixing with Bahasa Indonesia or local dialects, which may blur lexical boundaries that the model relies on for neutral-versus-polar disambiguation (Aji et al., 2022). Sundanese, while also exhibiting register distinctions (undak-usuk), has a less complex honorific tier system and may present cleaner sentiment signals in the NusaX-Senti corpus. We note that this interpretation is speculative, as GPT-4o-mini's closed-source nature precludes token-level error analysis. Future work employing open-source LLMs or confusion matrix inspection on larger samples would enable more rigorous linguistic attribution.

E-ISSN : 2830 – 7879 || P-ISSN : 2355 - 1887

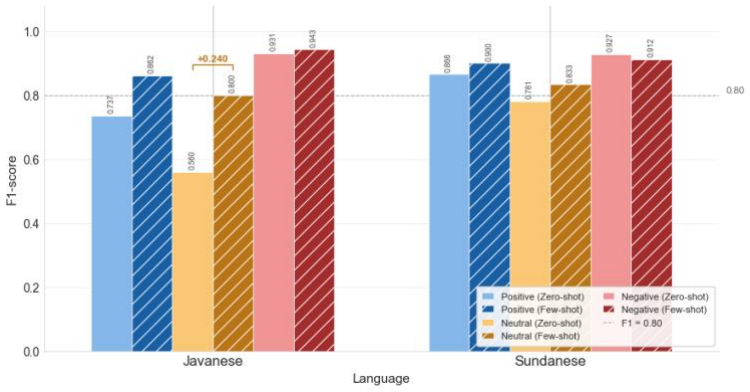

Figure 4. Per-class F1 scores

Figure 4. Per-class F1 scores for GPT-4o-mini under zero-shot (solid) and few-shot (hatched) prompting across Javanese and Sundanese (n=100 per language). The Neutral class exhibits the largest improvement from few-shot examples in Javanese (F1: 0.560 → 0.800, Δ=+0.240), explaining the cross-language performance gap observed under zero-shot prompting. The Negative class consistently achieves the highest F1 across all conditions, indicating clearer lexical polarity in both languages.

Across all GPT-4o-mini conditions, the Negative class consistently achieves the highest F1 (0.912–0.943 under few-shot), suggesting that negative sentiment carries the clearest lexical signal in both Javanese and Sundanese, a pattern consistent with multilingual SA findings reported in prior cross-lingual studies (Jazuli et al., 2024; Nababan & Alan, 2023).

Table 5. Per-class results for GPT-4o-mini (n=100 per language)

Language	Strategy	Class	Precision	Recall	F1
Javanese	Zero-shot	Positive	0.651	0.848	0.737
Javanese	Zero-shot	Neutral	0.737	0.452	0.560
Javanese	Zero-shot	Negative	0.919	0.944	0.931
Javanese	Zero-shot	Macro	0.769	0.748	0.743
Javanese	Few-shot	Positive	0.875	0.848	0.862
Javanese	Few-shot	Neutral	0.765	0.839	0.800
Javanese	Few-shot	Negative	0.971	0.917	0.943
Javanese	Few-shot	Macro	0.870	0.868	0.868
Sundanese	Zero-shot	Positive	0.853	0.879	0.866
Sundanese	Zero-shot	Neutral	0.758	0.806	0.781
Sundanese	Zero-shot	Negative	0.970	0.889	0.928
Sundanese	Zero-shot	Macro	0.860	0.858	0.858
Sundanese	Few-shot	Positive	1.000	0.818	0.900
Sundanese	Few-shot	Neutral	0.732	0.968	0.833
Sundanese	Few-shot	Negative	0.969	0.861	0.912
Sundanese	Few-shot	Macro	0.900	0.882	0.882

Table 6 shows IndoBERT-Base's per-class results and a clear prediction bias. The model assigns high precision to Neutral (1.000 for Javanese, 0.870 for Sundanese) but much lower recall (0.484 and 0.645). The positive and Negative F1 scores range from 0.061 to 0.150. Essentially, the model defaults to Neutral, identifying actual Neutral samples reasonably well when it does predict them, but it misses almost all Positive and Negative instances. This collapse to a conservative prediction strategy is typical of a classifier deployed outside its training distribution (Lossio-Ventura et al., 2024; Zouidine & Khalil, 2025).

Table 6. Per-class results for IndoBERT-Base

Language	Class	Precision	Recall	F1
Javanese	Positive	0.098	0.152	0.119

E-ISSN : 2830 – 7879 || P-ISSN : 2355 - 1887

Javanese	Neutral	1.000	0.484	0.652
Javanese	Negative	0.118	0.111	0.114
Javanese	Macro	0.405	0.249	0.295
Sundanese	Positive	0.128	0.182	0.150
Sundanese	Neutral	0.870	0.645	0.741
Sundanese	Negative	0.067	0.056	0.061
Sundanese	Macro	0.355	0.294	0.317

RQ3 answer: GPT-4o-mini's zero-shot performance is not uniform across languages. Sundanese leads by 0.112 Macro-F1 points, and the difference traces to the Neutral class in Javanese. Few-shot prompting closes the gap, so the cross-language difference is a prompting effect tied to Neutral class ambiguity in Javanese, not a general language-capability difference.

D. Statistical Significance (McNemar Test)

Table 7 summarizes all pairwise McNemar results. For RQ1, the GPT-4o-mini (few-shot) vs. IndoBERT-base comparison is highly significant on both languages: Javanese ($b=67$, $c=4$, $disc=71$, $\chi^2=54.14$, $p<0.001$) and Sundanese ($b=62$, $c=2$, $disc=64$, $\chi^2=54.39$, $p<0.001$). Discordant pair counts of 71 and 64 are well above the minimum threshold of 20. GPT-4o-mini's advantage is not a chance finding.

Table 7. McNemar test results

Comparison	Language	b	c	disc	χ^2	p-value	sig
GPT few-shot vs IndoBERT zero-shot	Javanese	67	4	71	54.14	<0.001	***
GPT few-shot vs IndoBERT zero-shot	Sundanese	62	2	64	54.39	<0.001	***
GPT few-shot vs GPT zero-shot	Javanese	14	3	17	5.88	0.015	* ⚠
GPT few-shot vs GPT zero-shot	Sundanese	5	3	8	0.13	0.724	ns ⚠

All tests use chi-squared with Yates' continuity correction. disc < 20 interpret with caution. For RQ2, the Javanese few-shot vs. zero-shot comparison gives $p=0.015$ but only 17 discordant pairs. Treated as suggestive, not definitive. The Sundanese comparison yields $p=0.724$ with 8 discordant pairs; not significant and underpowered. Both cases trace back to the small performance difference between prompting strategies, especially on Sundanese. Studies comparing prompting strategies should use $n > 200$ per language to achieve adequate power for modest effect sizes (Kocoń et al., 2023).

E. Comparison with Related Work and Practical Implications

GPT-4o-mini's best results (Macro-F1: 0.868-0.882) match or exceed published benchmarks for LLM-based sentiment analysis on other low-resource or non-English languages. GPT-4o on Arabic sentiment analysis yields 0.72-0.85 (Alahmadi, 2025; Rusnachenko et al., 2024) in-context learning for Spanish SA reaches roughly 0.75-0.82 (Nazi et al., 2025a). LLM benchmarks on standard Indonesian report 0.65-0.78 (Winata et al., 2023). With structured few-shot prompting, Javanese and Sundanese prove manageable for a modern LLM despite their low-resource status.

IndoBERT-base without NusaX-specific fine-tuning (0.295-0.317) lands well below fine-tuned BERT models in the literature, as expected. Domain-matched IndoBERT fine-tuning consistently yields Macro-F1 above 0.80 (Devlin et al., 2019; Lossio-Ventura et al., 2024). That gap reflects a mismatch in training distribution, not architecture. The practical takeaway: fine-tuned BERT

E-ISSN : 2830 – 7879 || P-ISSN : 2355 - 1887

remains strong when matched training data is available; GPT-4o-mini is a reasonable fallback when it is not.

For deployment, GPT-4o-mini with a few-shot prompt is the most practical configuration tested here. It needs no fine-tuning data, no GPU infrastructure, and only nine annotated examples. That is a substantially lower resource requirement than BERT-based fine-tuning, and it matters for Indonesian regional-language counts NLP where annotation budgets are tight. Three limitations are worth flagging. The sample size (n=100 per language) is sufficient for exploratory comparison, but limits statistical power for detecting small differences, particularly in McNemar tests for RQ2, as evidenced by the low discordant pair. Results are based on a single deterministic run (temperature=0.0), and variability across different prompt orderings was not evaluated. IndoBERT was tested without NusaX-specific fine-tuning, so the comparison does not reflect its best performance. Moreover, GPT-4o-mini's closed-source nature limits any linguistic-level explanation of the performance patterns. Future work should use larger sample sizes, include a NusaX fine-tuned IndoBERT baseline, extend evaluation to the remaining NusaX languages, and test chain-of-thought prompting as a third strategy.

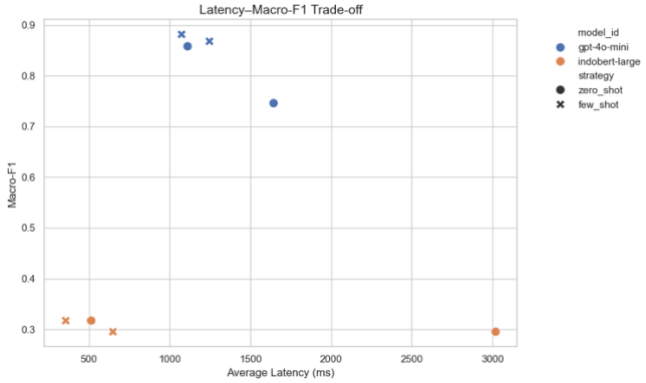

Figure 5. Latency-Macro-F1 trade-off across all experimental conditions

Figure 5. Latency-Macro-F1 trade-off across all experimental conditions (n=100 per language). Circle markers = zero-shot; cross markers = few-shot. GPT-4o-mini (blue) sits in the high-F1 / moderate-latency region; IndoBERT-base (orange) clusters in the low-F1 / low-latency region. The 3,020 ms IndoBERT zero-shot/Javanese outlier is API cold-start overhead. GPT-4o-mini API calls average 1,074-1,644 ms per inference; IndoBERT-base via HuggingFace ranges from 356-648 ms (few-shot) to 514-3,020 ms (zero-shot). GPT-4o-mini is slower but achieves Macro-F1 nearly three times higher. It is important to note that this is not a fully symmetrical comparison: GPT-4o-mini is evaluated in a zero-resource prompting scenario, while IndoBERT is deployed as a general-domain classifier without task-specific retraining. A fair architectural comparison would require fine-tuning IndoBERT directly on NusaX-Senti, which remains an important direction for future work.

F. Limitations and Scope

This study has several limitations that warrant consideration:

1. Sample size: With n=100 per language, statistical power is adequate for detecting large effects (RQ1: disc=71, 64) but limited for small prompting differences (RQ2: disc=17, 8). Future work should employ n≥200 per language to reliably detect modest effect sizes ($d<0.3$) with power ≥0.80.
2. Baseline configuration: IndoBERT-base was evaluated without NusaX-Senti-specific fine-tuning, representing a practical deployment scenario but not the model's architectural ceiling. The comparison assesses deployment readiness rather than maximum achievable performance. A domain-matched fine-tuned IndoBERT baseline remains essential for fair architectural comparison.

E-ISSN : 2830 – 7879 || P-ISSN : 2355 - 1887

3. Prompt design: Results reflect one prompt template and example ordering (seed=42). While deterministic sampling ensures reproducibility, alternative prompt designs-such as chain-of-thought formatting or different example sequences-may yield different performance. Prior work (OpenAI et al., 2023a) suggests few-shot classification is moderately robust to prompt variations, but systematic ablation studies would strengthen generalizability claims.
4. Language coverage: Evaluation covers 2 of 10 languages in NusaX-Senti. Generalization to other Indonesian regional languages (Minangkabau, Batak, Acehnese, etc.) requires empirical validation.
5. Closed-source model: GPT-4o-mini's proprietary nature limits linguistic-level analysis of why it performs better on Sundanese than Javanese in zero-shot settings. Open-source LLMs would enable deeper interpretability studies.
6. Cost and latency: While GPT-4o-mini requires no training infrastructure, API costs and inference latency (1-1.6 sec/sample) may limit scalability for large-scale production deployments compared to locally-hosted fine-tuned models.

Despite these limitations, our findings provide actionable insights for practitioners working with Indonesian regional languages where annotation budgets are constrained and rapid deployment is required.

4. CONCLUSION

This study compared GPT-4o-mini and IndoBERT-base on three-class sentiment analysis for two low-resource Indonesian regional languages, Javanese and Sundanese, using the NusaX-Senti benchmark (n=100 per language) under zero-shot and few-shot prompting with nine labeled examples per prompt. Findings should be interpreted as exploratory rather than definitive, given the sample size and the asymmetric comparison between a general-purpose LLM and a domain-mismatched classifier.

On RQ1, GPT-4o-mini outperforms the off-the-shelf IndoBERT model under a no-adaptation setting, highlighting its robustness when domain-specific fine-tuning data is unavailable. With few-shot prompting, it achieves Macro-F1 scores of 0.868 (Javanese) and 0.882 (Sundanese), compared to IndoBERT-Base's 0.295 and 0.317. McNemar tests confirm that the differences are not due to chance ($p<0.001$; discordant pairs: 71 and 64). Without any domain-specific fine-tuning, GPT-4o-mini performs substantially better under the evaluated conditions than off-the-shelf IndoBERT-base on both languages.

On RQ2, few-shot prompting improves GPT-4o-mini on Javanese at a statistically significant level ($\Delta F1=+0.122$, $p=0.015$), driven mainly by the Neutral class (F1: 0.560 to 0.800). On Sundanese, the gain is marginal and not significant ($\Delta F1=+0.024$, $p=0.724$), which fits the ceiling effect: the zero-shot baseline there (0.858) was already high. Whether few-shot examples help depends largely on how much room the model has to improve from zero-shot.

On RQ3, GPT-4o-mini's zero-shot performance differs across languages. Sundanese leads Javanese by 0.112 Macro-F1 points, and per-class analysis traces this to the Neutral class (Javanese: F1=0.560, Sundanese: F1=0.781). Few-shot prompting closes the gap, converging both languages to 0.868-0.882. The cross-language difference comes from Neutral-class ambiguity in Javanese, not from any broader resource-level difference between the two languages.

GPT-4o-mini with a few-shot prompt requires no training data, no fine-tuning infrastructure, and only 9 annotated examples, making it a viable option for Indonesian regional-language tasks where annotation is limited. Future work should test a NusaX-Senti fine-tuned IndoBERT baseline for a more direct architectural comparison, evaluate chain-of-thought prompting, extend to the remaining eight NusaX languages (including Minangkabau and Toba Batak), and use n>=200 per language to give the McNemar test adequate power for smaller effect sizes.

5. ACKNOWLEDGMENT

E-ISSN : 2830 – 7879 || P-ISSN : 2355 - 1887

The authors acknowledge the NusaX-Senti dataset creators (Winata et al., 2023) for providing publicly available benchmark data for Indonesian regional languages. This work was supported by the Department of Business Management and Information Technology, Universiti Muhammadiyah Malaysia, and the Department of Business and Informatics, Universitas Muhammadiyah Palangkaraya. We thank the reviewers for their constructive feedback that improved this manuscript.

REFERENCES

Aji, A. F., Winata, G. I., Koto, F., Cahyawijaya, S., Romadhony, A., Mahendra, R., Kurniawan, K., Moeljadi, D., Prasojo, R. E., Baldwin, T., Lau, J. H., & Ruder, S. (2022). One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7226–7249. https://doi.org/10.18653/v1/2022.acl-long.500

Alahmadi, D. (2025). Human Versus AI: A Comparative Study of Zero-Shot LLMs and Transformer Models Against Human Annotations for Arabic Sentiment Analysis. *International Journal of Advanced Computer Science and Applications*, *16*(8). https://doi.org/10.14569/IJACSA.2025.0160882

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020a). *Language Models are Few-Shot Learners* (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2005.14165

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020b). *Language Models are Few-Shot Learners* (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2005.14165

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., Sun, X., Li, L., & Sui, Z. (2024). A Survey on In-context Learning. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1107–1128. https://doi.org/10.18653/v1/2024.emnlp-main.64

Jazuli, A., Widowati, & Kusumaningrum, R. (2024). Optimizing Aspect-Based Sentiment Analysis Using BERT for Comprehensive Analysis of Indonesian Student Feedback. *Applied Sciences*, *15*(1), 172. https://doi.org/10.3390/app15010172

Kastrati, M., Imran, A. S., Hashmi, E., Kastrati, Z., Daudpota, S. M., & Biba, M. (2025). Unlocking language barriers: Assessing pre-trained large language models across multilingual tasks and unveiling the black box with Explainable Artificial Intelligence. *Engineering Applications of Artificial Intelligence*, *149*, 110136. https://doi.org/10.1016/j.engappai.2025.110136

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, Ł., Wojtasik, K., Woźniak, S., & Kazienko, P. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion*, *99*, 101861. https://doi.org/10.1016/j.inffus.2023.101861

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. *Proceedings of the 28th International Conference on Computational Linguistics*, 757–770. https://doi.org/10.18653/v1/2020.coling-main.66

Lossio-Ventura, J. A., Weger, R., Lee, A. Y., Guinee, E. P., Chung, J., Atlas, L., Linos, E., & Pereira, F. (2024). A Comparison of ChatGPT and Fine-Tuned Open Pre-Trained Transformers (OPT) Against Widely Used Sentiment Analysis Tools: Sentiment Analysis of COVID-19 Survey Data. *JMIR Mental Health*, *11*, e50150. https://doi.org/10.2196/50150

Nababan, A. M., & Alan, R. (2023). Aspect-Based Sentiment Analysis On Chatgpt In Twitter Using Bidirectional Encoder Representations From Transformers (Bert). *Journal of Theoretical and Applied Information Technology*, *101*(16), 6561–6568. Scopus.

Nasution, A. H., Onan, A., Murakami, Y., Monika, W., & Hanafiah, A. (2025). Benchmarking Open-Source Large Language Models for Sentiment and Emotion Classification in Indonesian Tweets. *IEEE Access*, *13*, 94009–94025. https://doi.org/10.1109/ACCESS.2025.3574629

Nazi, Z. A., Hossain, Md. R., & Mamun, F. A. (2025a). Evaluation of open and closed-source LLMs for low-resource language with zero-shot, few-shot, and chain-of-thought prompting. *Natural Language Processing Journal*, *10*, 100124. https://doi.org/10.1016/j.nlp.2024.100124

Nazi, Z. A., Hossain, Md. R., & Mamun, F. A. (2025b). Evaluation of open and closed-source LLMs for low-resource language with zero-shot, few-shot, and chain-of-thought prompting. *Natural Language Processing Journal*, *10*, 100124. https://doi.org/10.1016/j.nlp.2024.100124

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2023a). *GPT-4 Technical Report* (Version 6). arXiv. https://doi.org/10.48550/ARXIV.2303.08774

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2023b). *GPT-4 Technical Report* (Version 6). arXiv. https://doi.org/10.48550/ARXIV.2303.08774

Přibáň, P., Šmíd, J., Steinberger, J., & Mištera, A. (2024). A comparative study of cross-lingual sentiment analysis. *Expert Systems with Applications*, *247*, 123247. https://doi.org/10.1016/j.eswa.2024.123247

Rusnachenko, N., Golubev, A., & Loukachevitch, N. (2024). Large Language Models in Targeted Sentiment Analysis for Russian. *Lobachevskii Journal of Mathematics*, *45*(7), 3148–3158. https://doi.org/10.1134/S1995080224603758

Sinaga, F. M., Pangaribuan, J. J., -, K., -, F., & Widjaja, A. E. (2025). Dynamic Sentiment Analysis on the Emergence of Pre-Trained Generative Model-Based Applications in Indonesia. *International Journal of Advanced Computer Science and Applications*, *16*(12). https://doi.org/10.14569/IJACSA.2025.01612111

Suhartono, D., Wongso, W., & Tri Handoyo, A. (2024). IdSarcasm: Benchmarking and Evaluating Language Models for Indonesian Sarcasm Detection. *IEEE Access*, *12*, 87323–87332. https://doi.org/10.1109/ACCESS.2024.3416955

Wali, A., Alahmadi, A., Aljohani, A., & Alharbi, A. (2026). Evaluating Arabic Sentiment Analysis With GPT-4o: A Comparative Study of Raw and Preprocessed Text. *Journal of Cases on Information Technology*, *28*(1), 1–19. https://doi.org/10.4018/JCIT.402749

Winata, G. I., Aji, A. F., Cahyawijaya, S., Mahendra, R., Koto, F., Romadhony, A., Kurniawan, K., Moeljadi, D., Prasojo, R. E., Fung, P., Baldwin, T., Lau, J. H., Sennrich, R., & Ruder, S. (2023). NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 815–834. https://doi.org/10.18653/v1/2023.eacl-main.57

Wu, C., Ma, B., Zhang, Z., Deng, N., He, Y., & Xue, Y. (2025). Evaluating zero-shot multilingual aspect-based sentiment analysis with large language models. *International Journal of Machine Learning and Cybernetics*, *16*(10), 8079–8101. https://doi.org/10.1007/s13042-025-02711-z

Zouidine, M., & Khalil, M. (2025). Large Language Models for Arabic Sentiment Analysis and Machine Translation. *Engineering, Technology & Applied Science Research*, *15*(2), 20737–20742. https://doi.org/10.48084/etasr.9584

E-ISSN : 2830 – 7879 || P-ISSN : 2355 - 1887