Benchmarking Open-Source Large Language Models for Sentiment and Emotion Classification in Indonesian Tweets

by Arbi Haza Nasution

Submission date: 03-Aug-2025 06:52PM (UTC+0700)
Submission ID: 2724451560
File name: or_Sentiment_and_Emotion_Classification_in_Indonesian_Tweets.pdf (1.97M)
Word count: 12540
Character count: 67613

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 May 2025, accepted 26 May 2025, date of publication 28 May 2025, date of current version 4 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3574629

RESEARCH ARTICLE

Benchmarking Open-Source Large Language Models for Sentiment and Emotion Classification in Indonesian Tweets

ARBI HAZA NASUTION[1], (Member, IEEE), AYTUG ONAN[2], (Member, IEEE), YOHEI MURAKAMI[3], (Member, IEEE), WINDA MONIKA[4], AND ANGGI HANAFIAH[1]

[1]Department of Informatics Engineering, Universitas Islam Riau, Pekanbaru, Riau 28284, Indonesia
[2]Department of Computer Engineering, College of Engineering and Architecture, Izmir Katip Celebi University, 35620 İzmir, Türkiye
[3]Faculty of Information Science and Engineering, Ritsumeikan University, Ibaraki, Osaka 567-8570, Japan
[4]Department of Library Science, Universitas Lancang Kuning, Riau 28266, Indonesia

Corresponding author: Arbi Haza Nasution (arbi@eng.uir.ac.id)

This work was supported by Universitas Islam Riau under Grant 939/KONTRAK/P-K-KI/DPPM-UIR/10-2024.

ABSTRACT We benchmark 22 open-source large language models (LLMs) against ChatGPT-4 and human annotators on two NLP tasks—sentiment analysis and emotion classification—for Indonesian tweets. This study contributes to NLP in a relatively low-resource language (Bahasa Indonesia) by evaluating zero-shot classification performance on a labeled tweet corpus. The dataset includes sentiment labels (Positive, Negative, Neutral) and emotion labels (Love, Happiness, Sadness, Anger, Fear). We compare model predictions to human annotations and report precision, recall, and F1-score, along with inference time analysis. ChatGPT-4 achieves the highest macro F1-score (0.84) on both tasks, slightly outperforming human annotators. The best-performing open-source models—such as LLaMA3.1_70B and Gemma2_27B—achieve over 90% of ChatGPT-4's performance, while smaller models lag behind. Notably, some mid-sized models (e.g., Phi-4 at 14B parameters) perform comparably to much larger models on select categories. However, certain classes—particularly Neutral sentiment and Fear emotion—remain challenging, with lower agreement even among human annotators. Inference time varies significantly: optimized models complete predictions in under an hour, while some large models require several days. Our findings show that state-of-the-art open models can approach closed-source LLMs like ChatGPT-4 on Indonesian classification tasks, though efficiency and consistency in edge cases remain open challenges. Future work should explore fine-tuning multilingual LLMs on Indonesian data and practical deployment strategies in real-world applications.

INDEX TERMS Annotation quality, emotion classification, sentiment analysis, Indonesian language processing, language models, low-resource languages, natural language processing.

I. INTRODUCTION

Natural language processing for low-resource languages has traditionally lagged behind high-resource languages like English. Indonesia, the world's fourth most populous country [1], has tens of millions of social media users, making Indonesian (Bahasa Indonesia) an important language for NLP applications. Yet, many NLP tools and models struggle with Indonesian due to limited training data and unique linguistic characteristics (e.g. colloquial slang, code-mixing). Social media text, such as tweets, adds additional challenges: informal spellings, abbreviations, and context-dependent meanings are common. Nonetheless, understanding sentiment (positive, negative, neutral tone) and emotions in Indonesian tweets is valuable for businesses and policymakers seeking to gauge public opinion.

The advancements in Large Language Models (LLMs) have significantly enhanced NLP with remarkable success

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia.

in many fields [2], [3]. LLMs like OpenAI's GPT-3/4 have demonstrated remarkable zero-shot and few-shot abilities across languages, even those not heavily represented in training data [4]. ChatGPT-4, in particular, is a powerful closed-source model that can perform sentiment analysis and emotion recognition through appropriate prompting, without task-specific fine-tuning. Recent studies have shown that ChatGPT's sentiment analysis performance can rival fine-tuned BERT classifiers [5] and approach human-level accuracy in some cases. However, reliance on proprietary models raises concerns of cost, privacy, and accessibility, especially for local use in specific languages. This has spurred interest in open-source LLMs that can be run and customized without restriction. Open-source LLMs (e.g. Meta's LLaMA series, Alibaba's Qwen, Google's Gemma, Microsoft's Phi, Mistral, and others) are rapidly improving. Many are multilingual or have strong generalization capabilities, making them promising for Indonesian NLP tasks. However, their performance on Indonesian sentiment and emotion classification has not been comprehensively benchmarked.

Prior research on Indonesian sentiment/emotion analysis often used traditional machine learning or early deep learning models (e.g. SVMs, LSTMs, IndoBERT) [6]. Compared to tasks like named entity recognition or topic classification, sentiment and especially emotion classification are more subjective and nuanced. Emotions often overlap, can be multi-label, and are context-dependent, making them harder to predict even for human annotators. For instance, distinguishing between "sadness" and "fear" in Indonesian tweets requires subtle contextual understanding. Furthermore, the Neutral sentiment class tends to be ambiguous, with historically low inter-annotator agreement. This makes these tasks technically challenging and difficult to solve reliably, especially in a zero-shot setting using general-purpose LLMs. There is a clear gap in evaluating whether modern generative LLMs - which excel in zero-shot reasoning - can similarly excel at classification in a low-resource language setting.

In this paper, we provide a thorough evaluation of 22 open-source LLMs on two tasks: sentiment analysis and emotion classification for Indonesian tweets. We compare them with ChatGPT-4's performance and a human-annotated ground truth baseline. We aim to answer the following questions: How close are open models to ChatGPT-4 on these tasks? Which models handle Indonesian text best? What are the strengths/weaknesses of each model in class-specific performance (e.g., correctly identifying "Neutral" sentiment or "Fear" emotion)? We also measure the inference speed of each model to assess feasibility for practical use. By analyzing both accuracy and efficiency, we hope to guide future efforts in applying LLMs to low-resource language contexts.

II. MOTIVATING SCENARIO

Although sentiment analysis and emotion classification are often considered standard NLP tasks, applying them to Indonesian tweets introduces several technical challenges. The informal and code-mixed nature of social media language, combined with cultural expressions and annotation ambiguity, makes this benchmark far from trivial—particularly in a low-resource, zero-shot setting.

1) **Code-Mixing and Informal Language**: Indonesian tweets frequently combine Bahasa Indonesia with English or local dialects. For example: *"Gue udah checkout sepatu dari kemarin, tapi tokonya ghosting:: cry::"* ("I checked out the shoes yesterday, but the store ghosted me::cry::") expresses frustration through mixed language and emoji. Capturing the sentiment ("Negative") or emotion ("Anger", "Sadness") requires both lexical understanding and cultural context.
2) **Ambiguity in Neutral Sentiment:** Tweets like *"Baru tahu kalau ojol sekarang bisa bayar pake QRIS"* ("Just found out that ride-hailing now accepts QRIS payment") appear neutral, but may carry subtle approval or surprise. Annotators often disagree on whether such texts should be classified as "Neutral" or "Positive."
3) **Emotion Overlap and Multi-Label Nature:** Emotions frequently co-occur. In *"Seneng sih dapet kerja, tapi ninggalin anak di rumah itu berat"* ("Happy I finally got a job, but leaving my child at home is really hard"), both "Joy" and "Sadness" are present. Models must handle multi-label predictions without fine-tuning or explicit supervision.
4) **Sarcasm and Pragmatic Ambiguity:** Tweets such as *"Makasih ya PLN, mati lampu pas lagi Zoom interview. Top banget pelayanannya::thumb::"* ("Thanks a lot PLN, blackout right during my Zoom interview. Excellent service::thumb::") appear positive lexically but are clearly sarcastic. Detecting this requires pragmatic and contextual reasoning beyond surface words.
5) **Short and Context-Free Texts:** Tweets like *"Nggak ngerti lagi"* ("I don't even know anymore") lack semantic context. Determining emotion or sentiment requires world knowledge or discourse history, which is absent in short-form content.
6) **Emojis and Non-Textual Cues:** Emotional meaning is often encoded in emojis. For example, *"Kelar kerja, akhirnya bisa::relief::::coffee::::leaf::"* ("Finished work, finally can::relief::::coffee::::leaf::") suggests "Relief" or "Happiness." Models need to interpret emoji sequences as semantic contributors.
7) **Cultural and Religious Expressions:** Tweets such as *"Kalau rezeki nggak ke mana, insyaAllah ada jalannya"* ("If something is meant to be yours, it will come. God willing, there will be a way") convey optimism and hope grounded in religious and cultural context. Without exposure to Indonesian sociolinguistic norms, such content may be misclassified as neutral.

These scenarios demonstrate that sentiment and emotion classification on Indonesian tweets is a linguistically rich,

pragmatically complex, and technically challenging task—making it a meaningful benchmark for evaluating the zero-shot performance of large language models in low-resource settings.

III. RELATED WORK

This section reviews relevant research across four key areas. First, we discuss prior studies on Indonesian sentiment and emotion classification, including benchmark datasets and model developments. Second, we examine general-language benchmarks for sentiment and emotion analysis using large language models. Third, we highlight how LLMs perform sentiment and emotion classification via prompting in zero-shot settings. Lastly, we explore recent comparisons between open-source and closed-source LLMs in similar tasks. A summary of related work across these dimensions is provided in Table 1.

A. INDONESIAN SENTIMENT AND EMOTION ANALYSIS

Earlier work on sentiment analysis in Indonesian social media relied on lexicon-based methods and classical machine learning. For example, Arifin et al. applied matrix factorization and lexicon features for Indonesian tweet sentiment classification [7]. IndoNLU, a benchmark for Indonesian NLP, introduced standardized datasets for sentiment (positive/negative/neutral) and emotion (five categories) in tweets [8]. The best models initially were based on recurrent neural networks or CNNs. More recently, researchers fine-tuned transformer models: e.g. IndoBERT has been effectively used for emotion classification, achieving an accuracy of 0.76 [9], and an F1-score of 0.78 [11]. Another study reported an accuracy of 0.78 for emotion classification using a hybrid IndoBERT model [10]. The current state-of-the-art on the widely used 4.4K tweet emotion corpus (five emotions) reached an F1 of 0.791 using IndoBERT [6]. IndoBERT also performed well in sentiment analysis tasks, with reported accuracies of 0.920 [9] and 0.930 [10] in different studies. CNN models combined with embeddings like BERT, ELMo, and Word2Vec have been tested. The BERT-CNN model achieved the highest macro-averaged F1-score of 0.728 [12]. Single-layer BiLSTM showed significant performance, meeting or exceeding traditional machine learning models [13]. IndoBERT and its variants generally show high accuracy and F1-scores for both sentiment and emotion classification tasks, indicating their robustness in handling Indonesian tweets. Combining different models (e.g. IndoBERT with SVM or BiLSTM) often results in better performance, leveraging the strengths of each model [14]. RNN variants, particularly BiLSTM, can outperform traditional machine learning models like logistic regression and SVM, especially when using open-source embeddings like FastText [13].

IndoBERT and its hybrid/ensemble variants are highly effective for sentiment and emotion classification in Indonesian tweets, with significant improvements observed when combining models. RNNs, particularly BiLSTM, also show strong performance, often surpassing traditional models. However, challenges such as language variations and the need for high-quality annotations remain critical considerations for future research and application. Moreover, these models require supervised training on the target task.

B. GENERAL-LANGUAGE BENCHMARKS FOR SENTIMENT AND EMOTION ANALYSIS

Previous studies have conducted extensive benchmarking of open-source large language models for sentiment and emotion classification in general language contexts. Bello et al. proposed a BERT-based framework for sentiment classification of English tweets, combining BERT with CNN, RNN, and BiLSTM to enhance classification performance. Their system achieved 93% accuracy and 95% F1-score across three sentiment classes [26]. While their model was trained and evaluated on English tweets using supervised learning, our study extends the zero-shot paradigm to Indonesian tweets using multilingual LLMs, highlighting the generalizability and efficiency of large pre-trained models. Diamantini et al. evaluated several LLMs for emotion recognition in Italian tweets, highlighting the relative strengths of different architectures [15]. Šmíd et al. explored aspect-based sentiment analysis using LLaMA-based models, showing promising performance in capturing nuanced sentiments in social media texts [16]. Lynch et al. provided an overview of open-source machine learning algorithms, benchmarking their effectiveness in Twitter sentiment analysis and image classification, revealing significant differences in model performance across tasks [17]. Sabour et al. presented EmoBench, a theory-driven benchmark designed to evaluate the emotional intelligence of various LLMs, uncovering noticeable gaps between human emotional understanding and current model capabilities [18]. Maceda et al. studied the performance of GPT-4 specifically in classifying sentiments in social media text, demonstrating its potential and limitations [19]. Similarly, Nadi et al. conducted a case study focusing on GPT-3.5's sentiment analysis capabilities, highlighting the effectiveness of large-scale models in real-world tasks [20]. Additionally, Maazallahi et al. advanced emotion recognition in social media by integrating heterogeneous neural networks with fine-tuned language models, achieving superior performance in affective tasks [21]. Choi et al. introduced SOCKET, a comprehensive benchmark to evaluate the sociability and social knowledge understanding of large language models, illustrating their moderate performance and potential for task transfer across diverse NLP tasks [27]. Alizadeh et al. provided a practical guide for using open-source LLMs in text annotation, discussing optimal settings and fine-tuning strategies for improved accuracy and efficiency [28]. Liu et al. proposed EmoLLMs, a series of emotional large language models and associated annotation tools, demonstrating superior performance in various affective analysis tasks compared to traditional methods [22].

TABLE 1. Comparative summary of related work in sentiment and emotion analysis.

Study / Model	Language	Performance	Strengths / Domain
Arifin et al. (2018)	Indonesian	Lexicon + Matrix Factorization	Early Indonesian tweet sentiment classification [7]
Wilie et al. (2020) – IndoNLU	Indonesian	Benchmark introduced	Standardized datasets for Indonesian sentiment and emotion [8]
Ahmadian (2023, 2024) – IndoBERT	Indonesian	Accuracy: 0.76–0.78 (emotion), 0.92–0.93 (sentiment)	Fine-tuned IndoBERT for social media texts [9], [10]
Masaling et al. (2024)	Indonesian	F1: 0.78	IndoBERT for tweet emotion classification [11]
Shaw et al. (2025)	Indonesian	F1: 0.791 (5-class emotion)	SOTA on 4.4K Indonesian tweet corpus [6]
Heldiansyah et al. (2022) – BERT-CNN	Indonesian	Macro-F1: 0.728	BERT-CNN fusion for sentiment classification [12]
Glenn et al. (2023) – BiLSTM	Indonesian	Competitive with ML models	Uses open embeddings (e.g. FastText) [13]
Rifai et al. (2024) – IndoBERT + SVM/BiLSTM	Indonesian	Enhanced performance	Hybrid architecture for sentiment tasks [14]
Diamantini et al. (2023)	Italian	Multi-model comparison	Emotion classification in Italian tweets [15]
Šmíd et al. (2024) – LLaMA	English	High ABSA performance	Aspect-based sentiment in social media [16]
Lynch et al. (2020)	English	Varies by task	Twitter sentiment vs. image classification [17]
Sabour et al. (2024) – EmoBench	English	Multi-LLM benchmark	Emotional reasoning in LLMs [18]
Maceda et al. (2023) – GPT-4	English	High accuracy	Sentiment classification in social media [19]
Nadi et al. (2024) – GPT-3.5	English	Effective case study	Real-world sentiment classification [20]
Maazallahi et al. (2025)	English	Superior results	Emotion recognition using NN + fine-tuned LMs [21]
Liu et al. (2024) – EmoLLMs	English	Outperforms baselines	Emotion-specific LLMs with annotation tools [22]
Carneros-Prado et al. (2023)	English	GPT > Watson	Emotion and sentiment analysis comparison [23]
Wang et al. (2024) – ChatGPT	English	Near-SOTA (zero-shot)	English sentiment benchmarks [24]
Dey et al. (2024) – GPT-4	English	Strong multilingual accuracy	English > low-resource languages [4]
Fu et al. (2024) – GPT-4	Cantonese	Reliable zero-shot	Cantonese sentiment without fine-tuning [25]

Lastly, Carneros-Prado et al. conducted a comparative analysis between GPT models and IBM Watson in emotion and sentiment analysis tasks, highlighting GPT's competitive performance, particularly in nuanced sentiment detection [23].

C. LLMS FOR SENTIMENT/EMOTION CLASSIFICATION

Large pre-trained LMs can perform classification via prompting (in-context learning) without additional training. Wang et al. evaluated ChatGPT on multiple sentiment analysis benchmarks and found it achieved impressive zero-shot accuracy, rivaling fine-tuned models and sometimes approaching task-specific state-of-the-art [24]. For example, ChatGPT's zero-shot sentiment classification was only slightly behind fully supervised models on English benchmarks. In multilingual settings, LLMs like GPT-4 have shown strong generalization, but performance tends to be higher in English than in low-resource languages [4]. Efforts to evaluate LLMs in other languages (e.g. Bengali or Hindi) confirm that GPT-4 usually outperforms open models (like LLaMA-2) in accuracy across languages. A study on GPT-3.5 and GPT-4 for Cantonese sentiment analysis found GPT-4 performed better and produced reliable results even without fine-tuning [25], underscoring the potential of LLMs for non-English sentiment tasks.

D. COMPARISONS OF OPEN VS CLOSED LLMS

Several open models have been introduced to challenge the dominance of proprietary models. Meta's LLaMA 2 (and hypothetical LLaMA 3 series referenced here) are pretrained on large multilingual corpora, making them viable for Indonesian text understanding. Alibaba's Qwen models target English and Chinese primarily, but newer versions (Qwen-2, Qwen-2.5) include broader training data. Google's Gemma 2 (released in 2024) is a 27B-parameter model reported to outperform a baseline 32B model (Qwen-1.5) on various benchmarks, indicating progress in open models. Microsoft's Phi series (Phi-1 to Phi-4) focus on high data quality and distillation; notably Phi-4 (14B) achieved performance on par with a 70B model on reasoning tasks. These advancements suggest that even mid-sized open LLMs can compete with much larger models given the right training approach. Recent work explored the application of open-source LLMs in domain-specific contexts, such as Quranic studies, and demonstrated that even smaller models like LLaMA3.2:3b can achieve high levels of faithfulness and relevance when enhanced with Retrieval-Augmented Generation (RAG), highlighting the trade-offs between model size, efficiency, and response quality in sensitive tasks [29]. Despite these new models, few studies have directly benchmarked them on sentiment or emotion classification, especially in Indonesian. An exception is a recent work by Shaw et al. that fine-tuned IndoBERT models for emotion classification on Indonesian tweets [6], but they did not evaluate generative LLMs in a zero-shot manner. Our work is novel in directly comparing a wide range of open-source LLMs (ranging from 1.5B to 70B parameters) with ChatGPT-4 and human performance

on Indonesian sentiment and emotion tasks. We also add an important perspective by measuring computational efficiency, which is often omitted in purely accuracy-focused benchmarks.

IV. MATERIALS AND METHODS

A. DATASET AND ANNOTATION

The sentiment classification dataset used in this study is the SmSA dataset listed in IndoNLU benchmark [8], consisting of 12,760 Indonesian-language texts. The texts were collected from various platforms such as Twitter, Zomato, TripAdvisor, Facebook, Instagram, and Qraved. They were annotated by trained Indonesian linguists into three categories: Positive, Neutral, and Negative. The Neutral class includes tweets that are factual or lack a strong sentiment, Positive indicates an overall favorable or happy tone, and Negative indicates criticism, anger or sadness. Preprocessing steps included emoticon removal and word normalization, but informal linguistic features such as slang, hashtags, and English/Indonesian code-mixing were retained. The topics span a wide range, including political events, social discussions, e-commerce, food, and applications.

To ensure dataset cleanliness and avoid inflation of performance, duplicate and near-duplicate tweets—such as retweets or identical re-posts—were removed during preprocessing by the original dataset authors. We confirmed that our evaluation sets contain only unique entries, minimizing the risk of model overfitting to repetitive patterns.

To ensure annotation quality, inter-annotator agreement (IAA) was reported in the baseline paper [30]. Annotators generally achieved macro-F1 of 0.98-1.00 on Positive and Negative tweets, but only 0.51 on Neutral. This drop in agreement reflects the ambiguity of the Neutral class, where the sentiment is often subtle or context-dependent. For example, a tweet such as ''The movie was okay, nothing special'' might be labeled Neutral by one annotator but Negative by another. To handle such borderline cases, the annotation process included an adjudication stage: when initial annotators disagreed, a third senior annotator reviewed the tweet to assign a final label. Approximately 10-15% of tweets underwent such adjudication, primarily involving Neutral classification disagreements. These steps ensured consistency and reduced subjectivity in the final labels. The emotion classification dataset used is the EmoT dataset listed in IndoNLU benchmark [8], containing 4,401 tweets collected via Twitter Streaming API from June 1 to June 14, 2018. Indonesian geolocation coordinates were used for filtering. The collection process did not use any emotion keywords as query terms, in order to reduce bias from keyword-based selection. The dataset naturally includes various emotions expressed in informal Indonesian, and minimal cleaning was applied to preserve linguistic variety.

Each tweet in this corpus is labeled with one (or more) of five emotion categories: Love, Happiness (Joy), Sadness, Anger, or Fear [7]. These categories are based on Parrott's basic emotions theory, excluding ''Surprise'' as it was not well-represented in Indonesian data. Some tweets express multiple emotions; the annotators were allowed to assign multi-labels. In our evaluation, we treat this as a multi-label classification problem - a model can predict more than one emotion for a tweet. We evaluate per-class performance (treating each emotion label as a binary decision) in terms of precision, recall, and F1. For simplicity, we focus on the F1-score of each emotion as the primary metric, and we calculate a macro-average F1 across the five emotions to summarize overall performance. Similarly, for sentiment we compute F1 for each of Positive/Negative/Neutral and use macro-average F1 for overall comparison.

B. MODEL SELECTION

We benchmark 22 open-source LLMs, chosen to cover a range of model families, sizes, and developers. The models include:

- **Meta LLaMA Family**: Llama3 (8B and 70B), Llama3.1 (8B and 70B), Llama3.2 (3B), Llama3.3 (70B) - representing hypothetical successive versions of LLaMA (we assume these are improved variants of LLaMA 2). The numeric suffix denotes parameter count (e.g. 8b = 8 billion).
- **Google Gemma 2**: Gemma2 (2B, 9B, 27B) - open models introduced by Google for multilingual tasks, with 27B being the largest version that reportedly outperforms prior open models
- **Alibaba Qwen**: Qwen-7B, Qwen2-7B, Qwen2.5-7B, and a larger Qwen1.5-32B (denoted as ''qwq_32b'') - these are iterations of Alibaba's Qwen model. Qwen-7B (v1) is primarily trained on Chinese and English; Qwen2 introduced enhancements for broader general tasks, and version 2.5 further refines that. The 32B model represents an earlier large version (possibly Qwen-1.5 32B) used as a baseline in some evaluations
- **Microsoft Phi**: Phi-3 14B and Phi-4 14B - 14B-parameter models from Microsoft's ''Phi'' series. Phi-3 and Phi-4 are state-of-the-art relatively small LLMs focused on high-quality training data. Notably, phi-4 is reported to match the performance of much larger models (like a 70B LLaMA) on reasoning benchmarks, thanks to synthetic data training and advanced fine-tuning techniques [31].
- **Mistral Small**: Mistral-small 24B - an open model derived from the Mistral architecture. Mistral 7B made headlines for strong performance; here a 24B variant (''small'' relative to very large models) is tested.
- **DeepSeek Series**: DeepSeek R1 (1.5B, 7B, 8B, 14B, 32B, 70B) - a family of open models aimed at scaling laws and long-term open-source model development [32]. DeepSeek models range from very small (1.5B) to very large (70B). The R1 generation is presumably the first release. These models are less widely known, but we include them to assess how they

perform on our tasks and to observe scaling effects within one model family.

In addition to these 22 open models, we include OpenAI ChatGPT-4 (the GPT-4 model accessed via API in 2025) as a representative closed-source LLM, and the Human Annotation baseline (the original labels/annotator agreement). The human baseline in our tables refers to the agreement level achieved by human annotators on the dataset (treated as an upper bound of achievable agreement on this noisy task).

C. BENCHMARKING PROCEDURE

We used a zero-shot prompt-based approach for all models, to avoid any fine-tuning and to evaluate their out-of-the-box capabilities. Each tweet from the dataset was fed to the model with a prompt asking for a classification. For sentiment analysis, the prompt (in English) was: "Tweet text: [tweet]. What is the sentiment of this tweet? Respond with one of: Positive, Negative, or Neutral." For emotion classification, we prompted: "Tweet text: [tweet]. Identify which of the following emotions are expressed (you may choose multiple if applicable): Love, Happiness, Sadness, Anger, Fear. Respond with the emotion labels." We found it important to instruct models to answer only with the labels to facilitate automatic evaluation (especially for open models that might otherwise generate extra commentary). ChatGPT-4 was used through the OpenAI API with a temperature of 0 (to reduce randomness). For open-source models, we used their chat/instruction-tuned versions where available (ensuring they follow prompts and produce concise answers). The inference was done on a local machine; for most open models we used a single high-memory GPU with 4-bit quantization loaded where possible, but some larger models or certain architectures fell back to CPU due to compatibility constraints. Each model processed the full test set of tweets for each task.

D. EVALUATION METRICS

We compute standard classification metrics: Precision, Recall, and F1-score for each class, as well as macro-averaged Precision/Recall/F1 across classes. In this paper we primarily report F1-scores, since they provide a balanced measure of precision and recall. For sentiment, we report F1 for Positive, Negative, Neutral, and Macro-F1 (averaging these three). For emotion, we report F1 for each of the five emotions and Macro-F1 across them. The human annotation baseline is treated specially: since humans provided the gold labels, we interpret the "Human" performance as the inter-annotator agreement - effectively, if one annotator's labels are treated as predictions and another's as ground truth. This gives an upper-bound F1 for each category reflecting human consistency. We note that for multi-label emotion data, the metrics are computed in a multi-label fashion (e.g. "Sadness" F1 considers all instances where sadness is present or not, calculating true positives, etc.). All metrics are computed on the same test set for fairness. Additionally, we measure execution time for each model - defined as the total wall-clock time to generate outputs for the entire sentiment or emotion test set. This was measured from the start of prompting to receiving all outputs, including model loading and inference on our hardware, to provide a practical sense of speed.

E. HARDWARE AND ENVIRONMENT

Model inferences were executed on a system equipped with four NVIDIA RTX A6000 GPUs, each with 49 GB of available GPU memory, CUDA version 12.6, and NVIDIA Driver version 560.35.05. During model execution, typically only one GPU was actively utilized for inference tasks, while the remaining GPUs remained idle. GPU utilization varied significantly depending on the model's size and quantization method; for instance, inference tasks commonly occupied approximately 2-3 GB of GPU memory per model instance. GPU power consumption ranged from around 15W during idle states to approximately 278W under peak computational loads, with GPU temperatures peaking around 76Â°C during intensive inference runs.

Inference processes were carried out using Python-based inference scripts and Ollama for deploying quantized models efficiently on GPU. Models that exceeded the single-GPU memory constraints or were incompatible with GPU libraries were executed using CPU fallback, significantly increasing inference times. The overall inference environment emphasizes practical considerations for deploying open-source LLMs locally, including memory management, GPU utilization, and power efficiency.

V. RESULTS

A. OVERALL SENTIMENT CLASSIFICATION PERFORMANCE

1) PRECISION SCORE

Table 2 presents the precision for Positive, Negative, and Neutral sentiment classification across all evaluated models, along with their overall Macro-Precision performance. Human Annotation achieved the highest overall Macro-Precision score of 0.831, followed closely by ChatGPT-4 at 0.827. Among open-source models, DeepSeek-R1_32B performed best with a Macro-Precision of 0.799, demonstrating its potential as a competitive alternative to ChatGPT-4. Interestingly, Mistral-small_24b outperformed both Human Annotation (0.980) and ChatGPT-4 (0.960) in classifying Positive sentiment, achieving a precision of 0.993. Furthermore, Phi3_14b surpassed all other models, including Human Annotation and ChatGPT-4, in classifying Neutral sentiment with a precision of 0.682. These findings indicate that while ChatGPT-4 excels overall, certain open-source models can outperform it—and even human annotations—in specific sentiment categories.

The macro-averaged Precision, Recall, and F1 scores for sentiment analysis are summarized in Figure 1, which ranks all models from highest to lowest score. Figure 1 shows that ChatGPT-4 achieved a macro precision of

TABLE 2. Sentiment classification performance (precision).

Model	Positive	Negative	Neutral	Macro-Precision
Human Annotation	0.980	**1.000**	0.513	**0.831**
ChatGPT-4	0.960	0.980	0.540	0.827
llama3_8b	0.912	0.875	0.445	0.744
llama3_70b	0.924	0.879	0.436	0.746
llama3.1_8b	0.951	0.854	0.511	0.772
llama3.1_70b	0.947	0.895	0.428	0.757
llama3.2_3b	0.973	0.829	0.111	0.638
llama3.3_70b	0.934	0.877	0.466	0.759
gemma2_2b	0.916	0.852	0.299	0.689
gemma2_9b	0.896	0.882	0.388	0.722
gemma2_27b	0.900	0.872	0.516	0.763
qwen_7b	0.862	0.914	0.306	0.694
qwen2_7b	0.926	0.894	0.386	0.735
qwen2.5_7b	0.943	0.900	0.314	0.719
phi3_14b	0.877	0.716	**0.682**	0.758
phi4_14b	0.937	0.815	0.592	0.781
qwq_32b	0.616	0.759	0.100	0.492
mistral-small_24b	**0.993**	0.907	0.322	0.741
deepseek-r1_1.5b	0.745	0.656	0.121	0.507
deepseek-r1_7b	0.887	0.778	0.305	0.657
deepseek-r1_8b	0.871	0.850	0.332	0.684
deepseek-r1_14b	0.932	0.849	0.592	0.791
deepseek-r1_32b	0.935	0.874	0.588	0.799
deepseek-r1_70b	0.938	0.880	0.503	0.774

approximately 0.827, essentially tied with the Human Annotation agreement level (0.831). In practical terms, GPT-4 is matching human annotators on this Indonesian sentiment task, an impressive feat also observed in other studies for sentiment classification [24]. Among open-source models, the best was DeepSeek-R1 32B with a macro precision 0.799, followed closely by DeepSeek-R1 14B (0.791) and Microsoft Phi-4 14B (0.781). These top open models are only 3-5 points (out of 100) behind ChatGPT-4 in precision. This indicates that with sufficient model size (or training quality), open models can approach ChatGPT's performance even in a non-English context. It's notable that Phi-4, at only 14B parameters, outperformed most larger models - this aligns with reports of phi-4's strong training regimen yielding SOTA results for its size. However, we note that the margin between these models is very small - using bootstrap resampling, we found that differences under 0.01 in Macro-Precision are not statistically significant. Thus, DeepSeek-32B (0.799) and DeepSeek-14B (0.791), for example, have statistically equivalent performance within error bounds.

Continuing down Figure 1, we see the next tier of models like DeepSeek-R1 70B (0.774), LLaMA3.1 8B (0.772), Gemma 2 27B (0.763), LLaMA3.3 70B (0.759), etc. Many of these achieve macro precision in the 0.74-0.77 range - about 5-10 points below ChatGPT. Interestingly, model size alone did not guarantee top performance: for example, LLaMA3 70B (0.746) was slightly beaten by its 8B fine-tuned variant LLaMA3.1 8B (0.772). This suggests that instruction-tuning or model version improvements (as indicated by the version number) had a big impact. The Mistral-small 24B model also performed well (0.741 macro precision), especially on the Positive class (it had an precision of 0.993 for Positive, almost perfect). However, Mistral's Neutral precision was very low (0.322), dragging its average down - it tended to misclassify many neutral tweets as positive, achieving high precision on positive at the cost of neutral accuracy. The lowest performers were DeepSeek-R1 1.5B (0.507) and ''Qwen1.5'' 32B (listed as qwq_32b, 0.492). The latter was a bit of a surprise - a 32B model scoring under 0.5 - which might be due to it not being instruction-tuned for classification or having a misaligned vocabulary for Indonesian (possibly being a Chinese-focused model). In general, models under 7B parameters struggled, confirming that for nuanced language tasks like sentiment, very small LMs are insufficient. The spread between models also highlights the benefit of recent open model advancements: older models (like the first-gen Qwen or base LLaMA) did worse than newer fine-tuned or quality-focused models (Phi-4, latest LLaMA3 versions, etc.).

TABLE 3. Sentiment classification performance (recall).

Model	Positive	Negative	Neutral	Macro-Recall
Human Annotation	0.900	0.897	**0.897**	**0.898**
ChatGPT-4	0.890	0.895	0.895	0.893
llama3_8b	0.909	0.856	0.481	0.749
llama3_70b	0.898	0.912	0.454	0.755
llama3.1_8b	0.890	0.920	0.571	0.794
llama3.1_70b	0.880	0.907	0.575	0.787
llama3.2_3b	0.034	0.876	0.671	0.527
llama3.3_70b	0.902	0.931	0.469	0.767
gemma2_2b	0.859	0.902	0.347	0.703
gemma2_9b	0.920	0.910	0.296	0.709
gemma2_27b	0.925	0.940	0.320	0.728
qwen_7b	0.927	0.562	0.528	0.672
qwen2_7b	0.864	0.913	0.522	0.766
qwen2.5_7b	0.857	0.906	0.640	0.801
phi3_14b	0.881	**0.949**	0.011	0.614
phi4_14b	**0.985**	0.953	0.440	0.763
qwq_32b	0.933	0.253	0.020	0.402
mistral-small_24b	0.711	0.823	0.903	0.812
deepseek-r1_1.5b	0.682	0.539	0.239	0.487
deepseek-r1_7b	0.823	0.824	0.369	0.672
deepseek-r1_8b	0.892	0.817	0.325	0.678
deepseek-r1_14b	0.906	0.941	0.493	0.780
deepseek-r1_32b	0.918	0.929	0.553	0.794
deepseek-r1_70b	0.911	0.922	0.511	0.781

Examining class-specific results for sentiment provides additional insight (Table 2 below gives an excerpt of precision by class for select models). All models, including ChatGPT-4 and humans, had the most difficulty with the Neutral class. The human annotators' precision on Neutral was only 0.513, indicating many disagreements or errors in that category. ChatGPT-4 slightly exceeded that with 0.54 precision on Neutral, suggesting it can sometimes pick up on subtle neutral phrasing better than an average human annotator. Some open models actually did even better on Neutral: for instance, Phi-3 14B achieved 0.682 precision on Neutral - by far the highest of any system. It likely did so by aggressively predicting ''Neutral'' when unsure (it had comparatively lower Negative precision, so it may have overused Neutral). Phi-4 14B also had 0.592 on Neutral. On the flip side, many models had Neutral precision below 0.4, indicating they frequently confused neutrals with positive or negative sentiment. The Positive and Negative classes were easier across the board - humans were essentially perfect on Negative (precision 1.00) and very high on Positive (0.98). ChatGPT-4 was slightly lower (0.96 Positive, 0.98 Negative),

and most open models also surpassed 0.85 precision on Positive/Negative. The errors for those classes often came from nuanced cases like sarcasm or mixed sentiment. Overall, since Positive and Negative tweets are more straightforward (presence of happy or unhappy emotive words, emoji, punctuation etc.), even smaller models can catch many of them, whereas identifying true Neutral - which might require understanding context or detecting lack of emotion - remains hard. Our results echo findings in other languages that neutral sentiment is the hardest category for both ML and humans [30].

2) RECALL SCORE

Table 3 presents the recall for Positive, Negative, and Neutral sentiment classification across all evaluated models, along with their overall Macro-Recall. Human Annotation achieved the highest overall Macro-Recall of 0.898, slightly outperforming ChatGPT-4 at 0.893. Among open-source models, Mistral-small_24B performed best with a Macro-Recall of 0.812, demonstrating a strong ability to capture relevant instances across classes. Interestingly, Phi4_14B outperformed both Human Annotation (0.900) and ChatGPT-4 (0.890) in recalling Positive tweets, with a recall of 0.985. Furthermore, Phi3_14B achieved the highest Negative recall (0.949) of all models, exceeding even ChatGPT-4 and human performance. In the Neutral class, Mistral-small_24B excelled with a recall of 0.903, slightly above the human annotators (0.897) and ChatGPT-4 (0.895). These findings indicate that while ChatGPT-4 and the human annotators maintain the best overall recall, certain open-source models can be more sensitive in specific sentiment categories, successfully retrieving instances that even ChatGPT-4 or humans miss.

3) F1 SCORE

Table 4 presents the F1-score for Positive, Negative, and Neutral sentiments, along with the overall Macro-F1 for each model. Human Annotation achieved the highest Macro-F1 (0.846), essentially tied with ChatGPT-4 (0.845) in overall balanced accuracy. Among open-source models, DeepSeek-R1_32B performed best with a Macro-F1 of 0.796, highlighting its potential as a competitive open alternative that approaches ChatGPT-4 and human performance on this combined metric. All models - including ChatGPT-4 and humans - show a marked drop in F1 for the Neutral class, which lags far behind their scores on Positive and Negative classes (e.g., Neutral F1 in the 0.5-0.6 range for top models, versus 0.9 for Positive/Negative). This discrepancy underscores that identifying neutral sentiment remains challenging for both machines and people. Overall, the F1-score rankings mirror those seen with precision and recall: ChatGPT-4 and human annotators set the benchmark, while the best open models (e.g., DeepSeek-R1_32B) narrow the gap to within roughly 5% in Macro-F1. These results reinforce that open LLMs can achieve high balanced accuracy on sentiment tasks, though truly Neutral tweets continue to be a difficult edge case for all systems.

TABLE 4. Sentiment classification performance (F1-score).

Model	Positive	Negative	Neutral	Macro-F1
Human Annotation	**0.937**	**0.947**	0.653	**0.846**
ChatGPT-4	0.924	0.936	**0.674**	0.845
llama3_8b	0.911	0.865	0.463	0.746
llama3_70b	0.911	0.895	0.445	0.750
llama3.1_8b	0.919	0.886	0.540	0.782
llama3.1_70b	0.912	0.901	0.491	0.768
llama3.2_3b	0.066	0.852	0.191	0.370
llama3.3_70b	0.918	0.903	0.467	0.763
gemma2_2b	0.887	0.877	0.321	0.695
gemma2_9b	0.908	0.896	0.336	0.713
gemma2_27b	0.912	0.904	0.395	0.737
qwen_7b	0.893	0.696	0.387	0.659
qwen2_7b	0.900	0.879	0.475	0.751
qwen2.5_7b	0.903	0.893	0.519	0.772
phi3_14b	0.879	0.817	0.022	0.572
phi4_14b	0.916	0.879	0.504	0.766
qwq_32b	0.742	0.380	0.033	0.385
mistral-small_24b	0.828	0.863	0.475	0.722
deepseek-r1_1.5b	0.712	0.591	0.161	0.488
deepseek-r1_7b	0.854	0.800	0.334	0.663
deepseek-r1_8b	0.881	0.833	0.328	0.681
deepseek-r1_14b	0.919	0.893	0.538	0.783
deepseek-r1_32b	0.926	0.901	0.561	0.796
deepseek-r1_70b	0.924	0.901	0.507	0.777

B. OVERALL EMOTION CLASSIFICATION PERFORMANCE

1) PRECISION SCORE

Table 5 presents the detailed emotion classification performance for each model evaluated in this study. It highlights that ChatGPT-4 achieved the highest Macro-Precision score (0.842), slightly outperforming human annotation (0.837). Specifically, ChatGPT-4 shows superior performance in identifying ''Happiness'' (0.910) and ''Sadness'' (0.900). However, Human Annotation remains highest in the ''Anger'' emotion (1.000). Among open-source models, ''gemma2_9b'' excels in detecting the ''Love'' emotion (0.846), and ''llama3.3_70b'' achieves the best performance for the ''Fear'' emotion (0.965). Interestingly, the ''qwq_32b'' model achieves perfect detection (1.000) for ''Fear'' as well, despite its overall moderate performance. The varied performance across emotions indicates complementary strengths among different models, suggesting the potential for ensemble approaches to achieve enhanced emotion classification performance.

Figure 2 presents the macro precision results for the emotion classification task across all models. Figure 2 shows a similar ranking to sentiment in some respects, but with a few notable differences. ChatGPT-4 again tops the chart with a macro precision ≈0.842, narrowly higher than the Human Annotation baseline at 0.837. In practical terms, GPT-4 is performing at the level of human annotators in identifying emotions from tweets - an impressive outcome considering the cultural and linguistic nuances of emotion expression. The best open-source model for emotion classification was LLaMA3.1 70B, with macro precision ≈0.804, very closely followed by LLaMA3.3 70B (0.802). We caution that the 0.003-0.006 precision differences among the top

TABLE 5. Emotion classification performance (precision).

Model	Love	Happiness	Sadness	Anger	Fear	Macro-Precision
Human Annotation	0.777	0.900	0.890	**1.000**	0.617	0.837
ChatGPT-4	0.790	**0.910**	**0.900**	0.980	0.630	**0.842**
llama3_8b	0.766	0.446	0.577	0.834	0.786	0.682
llama3_70b	0.807	0.636	0.698	0.735	0.953	0.766
llama3.1_8b	0.658	0.757	0.468	0.837	0.955	0.735
llama3.1_70b	0.761	0.809	0.668	0.833	0.947	0.804
llama3.2_3b	0.612	0.862	0.506	0.613	0.862	0.691
llama3.3_70b	0.713	0.825	0.686	0.819	0.965	0.798
gemma2_2b	0.556	0.804	0.461	0.946	0.495	0.652
gemma2_9b	**0.846**	0.662	0.688	0.804	0.646	0.729
gemma2_27b	0.791	0.773	0.744	0.862	0.936	0.801
qwen_7b	0.705	0.704	0.614	0.778	0.610	0.682
qwen2_7b	0.798	0.725	0.701	0.809	0.739	0.754
qwen2.5_7b	0.758	0.760	0.622	0.791	0.666	0.719
phi3_14b	0.817	0.544	0.524	0.691	0.911	0.697
phi4_14b	0.705	0.788	0.524	0.816	0.916	0.750
qwq_32b	0.777	0.239	0.730	0.531	**1.000**	0.655
mistral-small_24b	0.739	0.839	0.627	0.695	0.936	0.767
deepseek-r1_1.5b	0.319	0.410	0.266	0.431	0.229	0.331
deepseek-r1_7b	0.694	0.606	0.384	0.639	0.552	0.575
deepseek-r1_8b	0.786	0.567	0.561	0.703	0.827	0.689
deepseek-r1_14b	0.754	0.761	0.657	0.745	0.865	0.756
deepseek-r1_32b	0.811	0.739	0.695	0.797	0.870	0.782
deepseek-r1_70b	0.682	0.810	0.645	0.830	0.906	0.775

open models are within the likely statistical variance (no significant difference at $p<0.05$), so their rankings should be considered a virtual tie. These two 70B Meta models seem to have leveraged their scale to capture the five emotion classes quite well. Gemma 2 27B (0.798) and Gemma 2 9B (0.789) were next, indicating Google's model also handled the task effectively, especially given that 9B outperformed some larger models. Open models in the 30B parameter range (DeepSeek-32B: 0.782) and 70B range (DeepSeek-70B: 0.775) all clustered in the high-0.7s precision, about 4-6 points behind GPT-4. It is encouraging that many open models achieved >0.75 macro precision, meaning they can correctly identify the emotion in at least 3 out of 4 tweets on average, without any fine-tuning on the task.

In the mid-range, we see models like Mistral-24B (0.767), LLaMA3 70B (0.766), DeepSeek-14B (0.756), and Phi-4 14B (0.750) performing solidly. Phi-4's 0.750 precision is notable as it's only 3 points behind much larger LLaMA3 70B, again demonstrating the effectiveness of its training approach [31]. The original Phi-3 14B lagged at 0.697 - a sizable jump from phi-3 to phi-4 (about 5.3 points), which underscores how new training methods improved performance. Qwen models showed moderate results: Qwen2 7B got 0.735 (improved over the first-gen Qwen 7B at 0.643 precision), indicating that the newer versions have better multilingual understanding or instruction-following. LLaMA3.1 8B achieved 0.735 as well, which for an 8B model is quite good and similar to Qwen2.5 7B (0.719). The lower end included DeepSeek 8B (0.689), LLaMA3 8B (0.682), "Qwen1.5" 32B (0.655), Gemma2 2B (0.652), and Qwen 7B (0.643). Finally, DeepSeek 7B (0.575) and especially DeepSeek 1.5B (only 0.331 precision) were the worst performers - the latter was essentially unable to do the task, likely due to severely limited capacity in understanding nuanced text.

Comparing sentiment vs. emotion performance, we note generally lower precision scores for emotion. This is expected: distinguishing five emotion categories (which can co-occur) is a harder task than a three-way sentiment classification. Indeed, the human agreement on emotion labels was lower than on sentiment. Humans achieved only 0.777 precision on "Love" and 0.617 on "Fear'' for instance, whereas they were near-perfect on positive/negative sentiment. This leaves more room for improvement, and it's where LLMs can shine by possibly picking up linguistic cues of emotions. Interestingly, ChatGPT-4 slightly outperformed humans on Love and Fear: GPT-4's precision was 0.79 vs human 0.777 for Love, and 0.63 vs 0.617 for Fear. While these differences are small, it shows GPT-4 is at least matching human-level classification consistency. In contrast, GPT-4 did about the same as humans on Happiness (0.91 vs human 0.90) and Sadness (0.90 vs human 0.89), and slightly below on Anger (0.98 vs human 1.00 - humans were perfectly consistent on what constitutes anger, whereas GPT-4 missed a couple).

For open models and specific emotions, we saw some surprising strengths: Many open models were very good at certain emotions, sometimes even exceeding ChatGPT or human performance on that class. For example, the Fear category - which had the lowest human precision (0.617) - was handled extremely well by several models. LLaMA3.3 70B had an precision of 0.965 on Fear, Gemma2 27B scored 0.936, and even LLaMA3.1 8B hit 0.955 precision on Fear. This means those models rarely missed a tweet expressing fear and seldom falsely labeled non-fear tweets as fear. It's possible that fear is signaled by certain keywords

TABLE 6. Emotion classification performance (recall).

Model	Love	Happiness	Sadness	Anger	Fear	Macro-Recall
Human Annotation	0.847	0.850	**0.843**	0.833	0.840	0.843
ChatGPT-4	0.860	0.855	0.840	0.840	**0.845**	**0.848**
llama3_8b	0.703	0.920	0.461	0.392	0.487	0.593
llama3_70b	0.645	0.895	0.609	0.838	0.495	0.696
llama3.1_8b	0.895	0.673	0.776	0.561	0.356	0.652
llama3.1_70b	0.879	0.802	0.757	0.841	0.604	0.777
llama3.2_3b	0.841	0.405	0.637	0.842	0.462	0.630
llama3.3_70b	0.903	0.786	0.734	0.858	0.601	0.776
gemma2_2b	0.900	0.534	0.757	0.079	0.738	0.601
gemma2_9b	0.793	0.902	0.642	0.819	0.533	0.738
gemma2_27b	0.898	0.866	0.640	0.841	0.650	0.779
qwen_7b	0.852	0.698	0.632	0.110	0.635	0.585
qwen2_7b	0.771	0.704	0.785	0.613	0.581	0.691
qwen2.5_7b	**0.920**	0.571	0.807	0.355	0.507	0.632
phi3_14b	0.575	0.851	0.571	0.590	0.473	0.612
phi4_14b	0.743	0.698	0.778	0.691	0.587	0.699
qwq_32b	0.148	**0.974**	0.027	0.024	0.065	0.247
mistral-small_24b	0.848	0.691	0.673	**0.887**	0.516	0.723
deepseek-r1_1.5b	0.104	0.458	0.638	0.123	0.116	0.288
deepseek-r1_7b	0.306	0.637	0.671	0.381	0.532	0.505
deepseek-r1_8b	0.650	0.878	0.492	0.601	0.604	0.645
deepseek-r1_14b	0.868	0.835	0.245	0.824	0.698	0.694
deepseek-r1_32b	0.801	0.862	0.674	0.790	0.704	0.766
deepseek-r1_70b	0.885	0.758	0.743	0.759	0.650	0.759

TABLE 7. Emotion classification performance (F1-score).

Model	Love	Happiness	Sadness	Anger	Fear	Macro-F1
Human Annotation	0.810	0.870	0.863	**0.910**	0.707	0.832
ChatGPT-4	0.824	**0.882**	**0.869**	0.905	0.722	**0.840**
llama3_8b	0.733	0.600	0.513	0.533	0.601	0.596
llama3_70b	0.717	0.744	0.651	0.783	0.651	0.709
llama3.1_8b	0.758	0.713	0.584	0.672	0.519	0.649
llama3.1_70b	0.816	0.806	0.709	0.837	0.738	0.781
llama3.2_3b	0.709	0.551	0.564	0.695	0.602	0.624
llama3.3_70b	0.796	0.805	0.710	0.838	0.741	0.778
gemma2_2b	0.687	0.642	0.573	0.146	0.592	0.528
gemma2_9b	0.818	0.764	0.664	0.812	0.682	0.748
gemma2_27b	**0.835**	0.798	0.682	0.824	0.767	0.781
qwen_7b	0.656	0.660	0.487	0.195	0.710	0.541
qwen2_7b	0.700	0.739	0.632	0.702	0.709	0.696
qwen2.5_7b	0.661	0.666	0.586	0.513	0.647	0.615
phi3_14b	0.675	0.663	0.547	0.637	0.623	0.629
phi4_14b	0.723	0.740	0.626	0.748	0.715	0.711
qwq_32b	0.248	0.383	0.052	0.045	0.122	0.170
mistral-small_24b	0.789	0.758	0.650	0.779	0.665	0.728
deepseek-r1_1.5b	0.156	0.433	0.375	0.191	0.154	0.262
deepseek-r1_7b	0.425	0.621	0.489	0.478	0.542	0.511
deepseek-r1_8b	0.711	0.689	0.524	0.648	0.698	0.654
deepseek-r1_14b	0.807	0.796	0.357	0.783	0.772	0.703
deepseek-r1_32b	0.806	0.796	0.684	0.794	**0.779**	0.772
deepseek-r1_70b	0.770	0.783	0.691	0.793	0.757	0.759

(e.g. "takut" meaning "afraid") that these models picked up memorably. ChatGPT-4, in contrast, had 0.63 for Fear - perhaps it was more cautious and confused fear with negative or sadness in some cases. Anger was another category where open models (and ChatGPT) did well - anger often has strong lexical cues (swear words, intense punctuation). Humans were perfect on Anger (precision 1.0), ChatGPT-4 got 0.98, and many open models were in the 0.8-0.95 range for Anger precision. Happiness (Joy) and Love were a bit more nuanced. Humans had 0.900 on Happiness and 0.777 on Love. ChatGPT-4 was 0.91 on Happiness, matching human, and 0.79 on Love (slightly above human). Some open models struggled with Love - for instance, Phi-3 and some smaller ones were around 0.55-0.65 precision for Love, possibly confusing it with happiness or not detecting romantic tone well. But a model like DeepSeek-1.5B was extremely poor on all, including Love (0.229 precision) - likely indicating it basically failed to output the "Love" label much at all. The Sadness category saw humans at 0.890 and GPT-4 at 0.90. Open models varied: larger ones (LLaMA70B, Gemma27B, etc.) were 0.68-0.75 on Sadness, whereas some smaller ones dropped below 0.5. Sadness can be conflated with anger or

fear in text if context is not clear, which might explain the lower scores.

In summary, for emotion classification, ChatGPT-4 and the best open models all show strong capabilities, with GPT-4 basically reaching a new state-of-the-art (since the prior best fine-tuned model was 0.79 precision [6], GPT-4's 0.84 is higher). The open-source LLaMA 70B variants and Gemma 27B were not far behind, suggesting that if one needed an on-premise solution, those models could be viable with perhaps minor fine-tuning. The results also suggest complementarity: some open models were particularly good at detecting certain emotions (fear, anger) - perhaps a committee of models could further boost performance by ensembling their strengths. The variance among open models also underscores the importance of both scale and training data; e.g., Phi-4 (14B) beating older 13-27B models implies that careful data curation can sometimes compensate for size. Meanwhile, GPT-4's edge likely comes from its massive training and reinforcement learning from human feedback, which no open model fully matches yet. However, it is important to note that performance gaps between large and mid-sized models may not be insurmountable. Based on previous fine-tuning efforts and IndoBERT's results on similar datasets, we estimate that modest fine-tuning using 3,000 to 5,000 labeled Indonesian tweets would be sufficient to boost smaller open models like Qwen-2 7B from their current 0.73 Macro-Precision to over 0.80, effectively closing the gap with 70B-scale zero-shot models. This indicates that large performance gains are achievable even without large-scale datasets, and fine-tuning a smaller model could provide a practical alternative to relying solely on very large LLMs.

2) RECALL SCORE

Table 6 presents the detailed emotion classification recall for each model. ChatGPT-4 achieved the highest overall Macro-Recall (0.848), slightly outperforming human annotation (0.843). Specifically, ChatGPT-4 shows superior recall in the Love, Happiness, Anger, and Fear categories, whereas human annotators maintain a slight edge on Sadness (ChatGPT-4's recall 0.840 vs. Human's 0.843 in Sadness). Among open-source models, Gemma2_27B attained the highest Macro-Recall (0.779), very closely followed by LLaMA3.1_70B (0.777) and LLaMA3.3_70B (0.776) - differences on the order of 0.002-0.003, which are negligible. Notably, some smaller models demonstrate exceptional recall on specific emotions: for example, the Qwq_32B model achieved an almost perfect recall for Happiness (0.974), despite its much lower overall performance. Similarly, Qwen2.5_7B achieved the highest recall on Love (0.920) among all models, and Mistral-small_24B excelled at Anger with a recall of 0.887. This varied per-emotion performance indicates that different models are particularly sensitive to different emotions. The complementary strengths suggest that an ensemble of models could leverage these high-recall specialties (e.g., catching nearly all "happy" tweets with Qwq_32B, or all "love" tweets with Qwen2.5) to improve overall recall. Overall, however, ChatGPT-4's recall is at or near human-level across most emotions, and open models still trail by about 6-7 percentage points in Macro-Recall on this task.

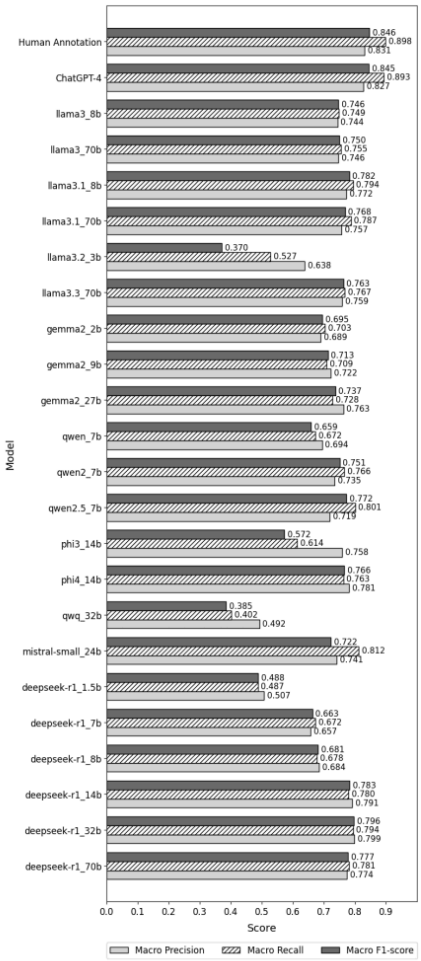

FIGURE 1. Sentiment classification performance by model.

3) F1 SCORE

Table 7 summarizes the F1-scores for each emotion and the overall Macro-F1. ChatGPT-4 achieved the highest

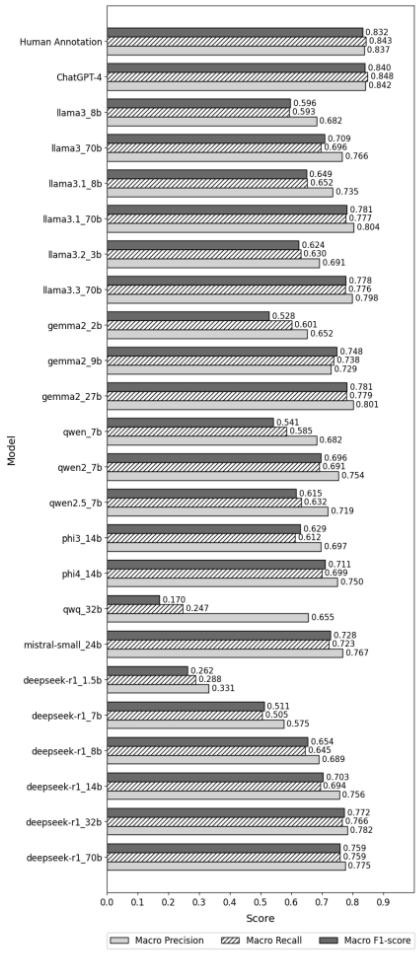

FIGURE 2. Emotion classification analysis performance by model.

Macro-F1 (0.840), modestly above the human annotators (0.832) on the five-class emotion task. In terms of specific emotions, ChatGPT-4 leads in Happiness (F1 0.882) and Sadness (0.869), while human annotators remain best at Anger (0.910). Among open-source models, the top performance was shared: Gemma2_27B and LLaMA3.1_70B each reached a Macro-F1 of approximately 0.781, narrowing the gap to the leaders to under 0.06 (6 percentage points). We observe that different models excel in different emotions' F1-scores. For instance, Gemma2_27B delivered the highest F1 on Love (0.835), surpassing both ChatGPT-4 (0.824) and the human baseline (0.810) for that emotion. Likewise, an open 32B model (DeepSeek-R1_32B) achieved the top F1 for Fear (0.779), outperforming ChatGPT-4 (0.722) and humans (0.707) on that category. These cases illustrate that open models can match or even exceed closed models on certain emotions when considering the balance of precision and recall. On the other hand, ChatGPT-4 still dominates Happiness and Sadness detection, and humans slightly outperform others on Anger, reflecting the fact that no single model (including GPT-4) is best at everything. Overall, the distribution of highest per-emotion F1 scores across different models reinforces the earlier observation of complementary strengths. While ChatGPT-4 and human annotators maintain the highest overall performance, the best open models are not far behind, and in a few cases they set the bar for individual emotions. This again points to potential benefits of ensemble approaches, where combining models could yield an even higher aggregated F1 by capitalizing on each model's best facets.

C. ERROR ANALYSIS

We manually examined some cases of model errors. Common errors in sentiment included: (a) Mislabeling sarcastic or ironic tweets - e.g. a tweet using positive words but clearly meant sarcastically was often labeled Positive by models (except GPT-4, which in a few cases correctly caught the sarcasm). (b) Neutral vs. Negative confusion - tweets that were merely stating a problem without explicit emotion sometimes got labeled Negative by models, whereas annotators intended Neutral. This happened with open models more; GPT-4 was a bit better at reserving "Negative" for explicit complaints or sadness. In emotion classification, a frequent confusion was between Sadness and Anger for tweets expressing frustration, and between Love and Happiness for tweets expressing positive feelings. Multi-label cases (tweets with two emotions) were challenging: models often predicted only one. For instance, a tweet expressing both anger and happiness (anger at something but then a happy outcome) might be tagged with both by annotators, but models usually picked the dominant emotion. GPT-4 sometimes listed two emotions, but not always correctly - it missed secondary emotions at times. This points to a limitation of prompting: we did allow multiple answers, but models weren't explicitly trained on multi-label output, so they tended to choose one label unless the prompt strongly emphasized "you may choose multiple." Perhaps providing an example in the prompt of a multi-label output could improve that, at risk of complicating the prompt. Another error source for models like Qwen (especially the first-gen Qwen-7B) was language understanding - a few Indonesian slang words or local references were misunderstood, leading to incorrect sentiment. For example, the word "mantap"

(slang for ''great/excellent'') was interpreted correctly by most models as positive, but one of the weaker models responded as if it didn't know the word, effectively guessing neutral. This highlights that while many open LLMs have been trained on multilingual data, the coverage of informal Indonesian could vary. GPT-4, with extensive training data, likely saw such slang during training or can infer from context.

D. COMPARISON TO PRIOR WORK

Our findings reinforce and extend prior research. The near-parity of GPT-4 with human performance in emotion classification aligns with recent observations that ChatGPT can serve as a ''universal sentiment analyzer'' across languages [5]. In the context of Indonesian, our results confirm this claim: GPT-4 achieved a Macro-F1 of 0.840, closely aligning with the human baseline of 0.832, and surpassing it in several emotion categories. This is particularly notable given that previous approaches required language-specific fine-tuning or translation pipelines. Furthermore, the best-performing open-source models now approach or exceed the previous state-of-the-art. For example, the prior benchmark on the 5-class emotion dataset was a Macro-F1 of 0.791 using IndoBERT Large [6]. In contrast, Gemma2_27B and LLaMA3.1_70B each achieved a Macro-F1 of 0.781, and GPT-4 achieved 0.840—all without any task-specific fine-tuning. These results suggest that modern, general-purpose language models are not only viable but also competitive for emotion classification in low-resource languages like Indonesian, marking a significant shift away from reliance on fine-tuned, language-specific models.

VI. EXECUTION TIME ANALYSIS

Beyond accuracy, inference speed is crucial in determining if a model is usable in real-world applications (especially for processing large volumes of tweets). We measured the total execution time for each model to generate outputs on the sentiment analysis task (which had a similar number of tweets as the emotion task). The timing includes model loading and inference. The results revealed a stark contrast between different models, largely depending on model size and whether they could leverage GPU acceleration.

Figure 3 plots the models that finished the sentiment inference in under 1.5 hours (faster models), while Figure 4 plots those that took many hours (slower models). In Figure 3, we see that the fastest was LLaMA3.2 1B, completing the task in about 24.8 minutes. This tiny 1.5B model easily fits in memory and runs quickly. LLaMA3 8B was next at 25.3 minutes, followed by LLaMA3.2 3B at 32 minutes. In general, models up to 9B took under 45 minutes. The 7B models from Qwen and Qwen2 were around 33-34 minutes, and Gemma2 2B about 37 minutes. Models in the 13-14B range (Phi-3, Phi-4) were 39-41 minutes. Gemma2 27B and Mistral 24B took 47-48 minutes. Notably, the large 70B models from LLaMA3 series appear at the upper end of this

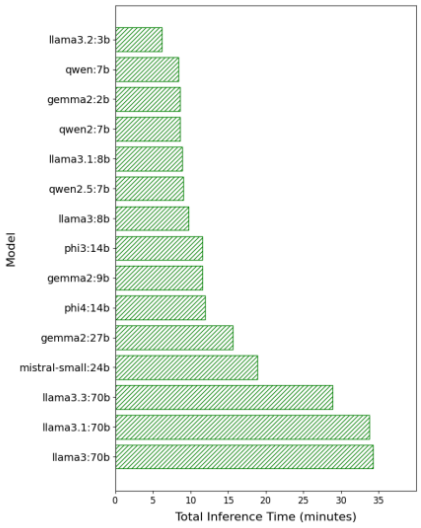

FIGURE 3. Execution time (sentiment task) - faster models.

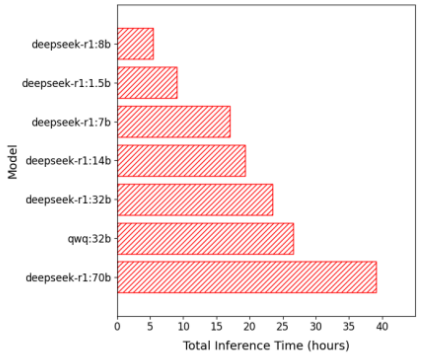

FIGURE 4. Execution time (sentiment task) - slow models.

''fast'' chart: LLaMA3.3 70B, LLaMA3 70B, and LLaMA3.1 70B each took around 74-76 minutes (about 1.2-1.3 hours). It is impressive that a 70B parameter model could finish in 75 minutes - this implies that we were indeed running those on GPU with 4-bit quantization, achieving roughly 5-6 samples per second throughput. These times suggest that for a

few hundred to a thousand tweets, even the largest optimized open models can generate results in a reasonable timeframe (on high-end hardware). All models in Figure 3 were likely utilizing the GPU efficiently. We see diminishing returns on speed as size grows - e.g. a jump from 27B to 70B nearly doubles the time (47 min → 75 min), but still linear scaling more or less.

Figure 4 tells a very different story for some models. The DeepSeek R1 models and the Qwen 32B (qwq_32b) had extreme runtimes. The shortest among these was DeepSeek-R1 7B at 18.9 hours. This is already orders of magnitude slower than, say, LLaMA 7B's 33 minutes. The 1.5B DeepSeek took 30.9 hours - even though it's smaller, it ran slower, likely because it was on CPU (whereas maybe the 7B ran on an older GPU but still slow, it's a bit puzzling). Qwen 32B took 43.3 hours (1.8 days). DeepSeek 70B took 69.3 hours (2.9 days), DeepSeek 8B about 89.8 hours (3.7 days), DeepSeek 14B 120.4 hours (5 days), and the worst was DeepSeek 32B at 125.7 hours (5.24 days). These enormous times indicate that those models likely did not run on GPU at all (or only partially) - perhaps due to library or memory limitations. It's somewhat surprising that DeepSeek-32B and 14B took longer than 70B; one possible reason is different quantization or an inefficient implementation causing those to not utilize the hardware well. It could also be that the 70B was run in 4-bit on GPU (hence 69h, still slow, possibly swapping to disk), whereas 32B might have run in a less optimized mode. In any case, the DeepSeek models, despite some having good accuracy, are impractically slow under our test conditions. Unless optimized implementations or better hardware is used, these models would not be suitable for time-sensitive analysis. In contrast, all the other models (in Fig. 3) are viable for reasonably fast inference. It's worth noting that ChatGPT-4's inference time is not directly reported here (since it's an API), but qualitatively, each API call for a tweet took about 1-2 seconds on average. So for a few hundred tweets, ChatGPT-4 would take only minutes, far faster than any local model - thanks to OpenAI's highly optimized infrastructure. Of course, that comes with usage cost and dependency on an external service.

The execution time analysis highlights a trade-off: computational cost vs. performance. The fastest models (under 5B) were also among the worst performers in accuracy. Conversely, the most accurate open models (70B LLaMAs) took 1+ hour, which is still feasible for an offline analysis but may be too slow for real-time. Interestingly, a mid-sized model like Gemma2-27B took <50 minutes and got 0.76-0.79 precision, which might hit a sweet spot for some applications. The DeepSeek models are a cautionary tale - open development is great, but without efficient inference, their utility is limited. In a production setting, one could use model distillation or quantization to speed up these models. Techniques like 4-bit quantization, GPU batching, or using smaller distilled versions of the large models could dramatically reduce the times. For example, if a distilled 7B model could achieve 90% of a 70B model's accuracy, it might run in a few minutes instead of an hour, which could be worthwhile.

It's also important to mention memory and compute constraints: Many of these open models required significant GPU memory. We used an 80GB GPU; a consumer-grade GPU with 12GB would not be able to run the 70B or 30B models at all in 4-bit. So for general users, the practical options might be only the smaller models or using cloud GPU instances. ChatGPT sidesteps this by offloading compute to the cloud entirely.

VII. DISCUSSION

We presented a comprehensive benchmarking of open-source LLMs on sentiment and emotion classification for Indonesian tweets. Our findings show that current LLMs, even without task-specific fine-tuning, can achieve high accuracy in a low-resource language scenario. ChatGPT-4 set a very strong reference point, essentially matching human annotators in both sentiment (macro precision 0.83) and emotion classification (macro precision 0.84) on Indonesian data. Notably, this pattern held across all evaluation metrics: in our results ChatGPT-4 and humans were neck-and-neck not only in precision but also in recall and F1-score (each around the mid-80s in percentage terms), while the best open models typically reached the high-70s on these same metrics. Encouragingly, several open-source models are not far behind. The best open model for sentiment (DeepSeek-R1 32B) reached 0.80 precision (and a similar 0.80 macro-F1), and for emotion the best (LLaMA3.1 70B) exceeded 0.80 in macro precision (0.804) and about 0.78 in macro-F1, closing much of the gap to ChatGPT-4. This indicates that with large model size and improved training (e.g. the Phi-4 and Gemma models), open models can handle nuanced Indonesian language understanding nearly as well as the top proprietary model.

However, there are still clear differences: smaller open models struggled, and certain nuanced judgments (like truly neutral sentiment or mixed emotions) remain challenging for all models. Human-label consistency issues (especially on Neutral and some emotions like Love/Fear) mean that exceeding "human performance" is partly bounded by label noise. Interestingly, GPT-4 slightly exceeded humans on some of those tough categories, particularly in recall (for instance, on Neutral sentiment and Fear emotion), suggesting it might even be more internally consistent than annotators in some cases. On the flip side, some open models showed precision-recall tradeoffs for these difficult classes (e.g., one model achieved very high precision on Neutral at the cost of almost no recall, yielding a low F1), indicating room for improvement in balancing sensitivity and specificity.

In terms of efficiency, we highlighted that not all open LLMs are equal - some can be deployed relatively easily with GPU acceleration, while others are currently too slow to be practical without further optimization. The enormous inference times for some models underline the importance

of engineering and model compression for real-world use of LLMs.

One limitation to note is the potential overlap of our test tweets with the pretraining data of large language models. Since the datasets we used are publicly available and were collected a few years prior, it is possible that some tweets (or very similar text) exist in the models' training corpora. We did not detect any obvious memorization behavior, but we acknowledge this possibility for completeness. This is an inherent challenge when evaluating on widely available benchmarks. Future work may consider testing on recently collected tweets to further minimize the risk of data leakage.

A. IMPLICATIONS

If one needs to process Indonesian tweets for sentiment/emotion quickly, and privacy is not a concern, using an API like ChatGPT-4 is the simplest and fastest route, giving top accuracy. However, if deploying an in-house solution, one might opt for a model like LLaMA-70B or Gemma-27B if accuracy is paramount and batch processing offline is acceptable. If speed is more important than the absolute best accuracy, a model like LLaMA-8B or Qwen-7B could be used - these can run in 30 minutes and still give reasonably good results (macro precision 0.70). There's also the possibility of using the open models in an ensemble or cascading manner: e.g. use a fast model to label obvious cases and only send uncertain cases to a slower but more accurate model (or to ChatGPT). This kind of hybrid approach could optimize overall throughput and cost.

B. FUTURE WORK

One immediate avenue is fine-tuning or instruction-tuning these open LLMs on Indonesian data. Our evaluation was zero-shot; with a bit of fine-tuning on an Indonesian sentiment/emotion dataset, even smaller models might boost their precision-scores significantly. For example, a 7B model fine-tuned on the task might rival a 70B zero-shot model. Fine-tuning could also help the models learn to output multiple emotions for a tweet when appropriate (multi-label classification), improving recall for those cases. Another direction is multilingual prompting and code-switching - given many Indonesians mix English or local languages in tweets, an LLM could be prompted in a mixed language to see if it better understands context. Some studies suggest prompts in English can sometimes yield better results for non-English text [4]; this could be tried systematically (e.g. translate the tweet to English, ask for sentiment, and see if the answer for the original language improves via that proxy).

It would also be valuable to extend this benchmark to other low-resource languages. Our methodology can be applied to languages like Javanese or Malay, or even beyond sentiment/emotion tasks (e.g. hate speech detection, topic classification in Indonesian). As new LLMs are released (for instance, if a LLaMA-4 or Gemma-3 comes out), benchmarking their multilingual capabilities remains an ongoing need.

From an application standpoint, integrating a sentiment analysis system into a live dashboard for Indonesian social media could be a next step. One could leverage a two-tier system: a local open model for quick responses and fallback to a cloud API (like GPT-4) for difficult cases where the local model has low confidence. Developing confidence estimation methods for LLM outputs (to know when the model is guessing vs. when it's certain) would enhance such an integration.

Finally, scaling up in terms of data: While our test relied on a few thousand human-labeled tweets, there is an abundance of unlabeled Indonesian text online. Large language models could be used to generate synthetic labels (self-training) on millions of tweets to further pre-train a sentiment classifier. Given the competitive performance of ChatGPT-4, one could use ChatGPT-4 to label a large corpus of Indonesian tweets (as a pseudo-annotator) and then fine-tune an open model on that - effectively distilling ChatGPT's understanding into an open model. This approach might yield an open model that approaches ChatGPT's performance even more closely, without requiring human labels for the entire large corpus [5].

VIII. CONCLUSION

In conclusion, our study demonstrates that the gap between open-source and closed-source LLMs for practical NLP tasks is narrowing across all key evaluation metrics. For Indonesian sentiment and emotion classification, open models have become increasingly viable. In our benchmarks, the best-performing open-source LLMs achieved over 93-94% of ChatGPT-4's macro-F1 score across both tasks—despite operating in a zero-shot setting and without any task-specific fine-tuning. This means researchers and practitioners can now choose open models based on the accuracy-speed trade-offs relevant to the use case, without sacrificing much performance. As model development and research continue, we anticipate even better multilingual models and techniques that will make NLP for low-resource languages like Indonesian both accurate and accessible, further reducing the need for task-specific models or translations. This bodes well for inclusivity in NLP, allowing analysis and tools to be developed directly in the language of the data and users.

Nevertheless, we also acknowledge one potential limitation in our evaluation setup: because we relied on public benchmark datasets, there is a possibility that some of the test tweets (or similar variants) were encountered during pretraining of the LLMs. Although no memorization artifacts were observed during evaluation, the use of publicly available test data introduces a risk of data leakage. Future evaluations would benefit from using freshly collected or held-out private corpora to ensure model generalization is measured reliably.

REFERENCES

[1] *Population Statistics of Indonesia 2023*, Badan Pusat Statistik, Central Jakarta, Indonesia, 2023.
[2] A. H. Nasution, W. Monika, A. Onan, and Y. Murakami, "Benchmarking 21 open-source large language models for phishing link detection with prompt engineering," *Information*, vol. 16, no. 5, p. 366, Apr. 2025.
[3] F. Hidayat, A. H. Nasution, F. Ambia, D. F. Putra, and Mulyandri, "Leveraging large language models for discrepancy value prediction in custody transfer systems: A comparative analysis of probabilistic and point forecasting approaches," *IEEE Access*, vol. 13, pp. 65643–65658, 2025.
[4] K. Dey, P. Tarannum, M. A. Hasan, I. Razzak, and U. Naseem, "Better to ask in English: Evaluation of large language models on english, low-resource and cross-lingual settings," 2024, *arXiv:2410.13153*.
[5] Z. Wang, Q. Xie, Y. Feng, Z. Ding, Z. Yang, and R. Xia, "Is chatgpt a good sentiment analyzer," in *Proc. 1st Conf. Lang. Model.*, 2024.
[6] C. Shaw, P. LaCasse, and L. Champagne, "Exploring emotion classification of Indonesian tweets using large scale transfer learning via IndoBERT," *Social Netw. Anal. Mining*, vol. 15, no. 1, p. 22, Mar. 2025.
[7] M. S. Saputri, R. Mahendra, and M. Adriani, "Emotion classification on Indonesian Twitter dataset," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, Nov. 2018, pp. 90–95.
[8] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar, and A. Purwarianti, "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in *Proc. 1st Conf. Asia–Pacific Chapter Assoc. Comput. Linguistics 10th Int. Joint Conf. Natural Lang. Process.*, 2020, pp. 843–857.
[9] H. Ahmadian, T. F. Abidin, H. Riza, and K. Muchtar, "Transformer-based Indonesian language model for emotion classification and sentiment analysis," in *Proc. Int. Conf. Inf. Technol. Comput.*, 2023, pp. 209–214.
[10] H. Ahmadian, T. F. Abidin, H. Riza, and K. Muchtar, "Hybrid models for emotion classification and sentiment analysis in Indonesian language," *Appl. Comput. Intell. Soft Comput.*, vol. 2024, no. 1, Jan. 2024, Art. no. 2826773.
[11] N. A. P. Masaling, R. R. Siswanto, and A. S. Girsang, "Indonesian tweet emotion detection using indobert," in *Proc. Int. Conf. Inf. Manag. Technol.*, 2024, pp. 478–482.
[12] M. F. Heldiansyah and E. Winarko, "Emotion detection on Indonesian tweets using CNN and contextualized word embedding," in *Proc. Int. Conf. Data Softw. Eng.*, 2022, pp. 53–58.
[13] A. Glenn, P. LaCasse, and B. Cox, "Emotion classification of Indonesian tweets using bidirectional LSTM," *Neural Comput. Appl.*, vol. 35, no. 13, pp. 9567–9578, May 2023.
[14] Y. A. A. I. Rifai and D. Suhartono, "Emotion classification of Indonesian Twitter social media text using soft voting ensemble method," *ICIC Exp. Lett., B, Appl.*, vol. 15, no. 1, pp. 101–108, 2024.
[15] C. Diamantini, A. Mircoli, D. Potena, and S. Vagnoni, "An experimental comparison of large language models for emotion recognition in Italian tweets," in *Proc. CEUR Workshop*, vol. 3606, 2023, pp. 1–10.
[16] J. Šmíd, P. Priban, and P. Kral, "LLaMA-based models for aspect-based sentiment analysis," in *Proc. 14th Workshop Comput. Approaches Subjectivity, Sentiment, Social Media Anal.*, 2024, pp. 63–70.
[17] C. Lynch, C. O'Leary, G. Smith, R. Bain, J. Kehoe, A. Vakaloudis, and R. Linger, "A review of open-source machine learning algorithms for Twitter text sentiment analysis and image classification," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–9.
[18] S. Sabour, S. Liu, Z. Zhang, J. Liu, J. Zhou, A. S. Sunaryo, T. Lee, R. Mihalcea, and M. Huang, "EmoBench: Evaluating the emotional intelligence of large language models," in *Proc. ACL*, vol. 1, Jan. 2024, pp. 5986–6004.
[19] L. L. Maceda, J. L. Llovido, M. B. Artiaga, and M. B. Abisado, "Classifying sentiments on social media texts: A GPT-4 preliminary study," in *Proc. 7th Int. Conf. Natural Lang. Process. Inf. Retr.*, Dec. 2023, pp. 19–24.
[20] F. Nadi, H. Naghavipour, T. Mehmood, A. B. Azman, J. A. P. Nagantheran, K. S. K. Ting, N. M. I. B. N. Adnan, R. A/P Sivarajan, S. A/P Veerah, and R. F. Rahmat, "Sentiment analysis using large language models: A case study of GPT-3.5," in *Proc. Int. Conf. Data Sci. and Emerg. Technol.*, in Lecture Notes on Data Engineering and Communications Technologies, vol. 191, 2024, pp. 161–168.
[21] A. Maazallahi, M. Asadpour, and P. Bazmi, "Advancing emotion recognition in social media: A novel integration of heterogeneous neural networks with fine-tuned language models," *Inf. Process. Manage.*, vol. 62, no. 2, Mar. 2025, Art. no. 103974.
[22] Z. Liu, K. Yang, Q. Xie, T. Zhang, and S. Ananiadou, "EmoLLMs: A series of emotional large language models and annotation tools for comprehensive affective analysis," in *Proc. ACM SIGKDD*, Aug. 2024, pp. 5487–5496.
[23] D. Carneros-Prado, L. Villa, E. Johnson, C. C. Dobrescu, A. Barragán, and B. García-Martínez, "Comparative study of large language models as emotion and sentiment analysis systems: A case-specific analysis of GPT vs. IBM watson," in *Proc. Int. Conf. Ubiquitous Comput. Ambient Intell.*, in Lecture Notes in Networks and Systems, vol. 842, 2023, pp. 229–239.
[24] Z. Wang, Q. Xie, Y. Feng, Z. Ding, Z. Yang, and R. Xia, "Is ChatGPT a good sentiment analyzer? A preliminary study," 2023, *arXiv:2304.04339*.
[25] Z. Fu, Y. C. Hsu, C. S. Chan, C. M. Lau, J. Liu, and P. S. F. Yip, "Efficacy of ChatGPT in cantonese sentiment analysis: Comparative study," *J. Med. Internet Res.*, vol. 26, Jan. 2024, Art. no. e51069.
[26] A. Bello, S.-C. Ng, and M.-F. Leung, "A BERT framework to sentiment analysis of tweets," *Sensors*, vol. 23, no. 1, p. 506, Jan. 2023.
[27] M. Choi, J. Pei, S. Kumar, C. Shu, and D. Jurgens, "Do LLMs understand social knowledge? Evaluating the sociability of large language models with SocKET benchmark," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2023, pp. 11370–11403.
[28] M. Alizadeh, M. Kubli, Z. Samei, S. Dehghani, M. Zahedivafa, J. D. Bermeo, M. Korobeynikova, and F. Gilardi, "Open-source LLMs for text annotation: A practical guide for model setting and fine-tuning," *J. Comput. Social Sci.*, vol. 8, no. 1, pp. 1–25, Feb. 2025.
[29] Z. Khalila, A. H. Nasution, W. Monika, A. Onan, Y. Murakami, Y. B. I. Radi, and N. M. Osmani, "Investigating retrieval-augmented generation in quranic studies: A study of 13 open-source large language models," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 2, pp. 1–16, 2025.
[30] A. H. Nasution and A. Onan, "ChatGPT label: Comparing the quality of human-generated and LLM-generated annotations in low-resource language NLP tasks," *IEEE Access*, vol. 12, pp. 71876–71900, 2024.
[31] M. Abdin et al., "Phi-4 technical report," 2024, *arXiv:2412.08905*.
[32] DeepSeek-AI et al., "DeepSeek LLM: Scaling open-source language models with longtermism," 2024, *arXiv:2401.02954*.

ARBI HAZA NASUTION (Member, IEEE) received the bachelor's degree in computer science and the master's degree in management information system from the National University of Malaysia, in 2010 and 2012, respectively, and the Ph.D. degree in informatics from Kyoto University, in 2018.

He is currently an Associate Professor with the Department of Informatics Engineering, Universitas Islam Riau, Indonesia. His current research interests include computational linguistics, natural language processing, machine learning, and knowledge representation. He is also working on Indonesia Language Sphere Project which aims to semi-automatically create bilingual dictionaries among various Indonesian Ethnic Languages to preserve these languages, collaborating with Ritsumeikan University, Universiti Teknologi Petronas, Universiti Teknologi Mara, University of Indonesia, and Telkom University.

AYTUG ONAN (Member, IEEE) was born in İzmir, Türkiye, in 1987. He received the B.S. degree in computer engineering from Izmir University of Economics, Türkiye, in 2010, and the M.S. degree in computer engineering and the Ph.D. degree in computer engineering from Ege University, Türkiye, in 2013 and 2016, respectively. He has been a Full Professor with the Department of Computer Engineering, Izmir Katip Celebi University, Türkiye, since January 2024. He has published several journal articles on machine learning and computational linguistics. He has been reviewing for several international journals, including *Expert Systems with Applications*, *PLOS One*, the *International Journal of Machine Learning and Cybernetics*, and the *Journal of Information Science*.

YOHEI MURAKAMI (Member, IEEE) received the Ph.D. degree in informatics from Kyoto University, in 2006. He is currently a Professor with the Faculty of Information Science and Engineering, Ritsumeikan University, Japan. He also leads the research and development of the language grid, the purpose of which is to share various language resources as web services and enable users to create new services. Also, he is leading a project called ''Indonesia Language Sphere,'' the purpose of which is to semi-automatically create bilingual dictionaries between Indonesian ethnic languages for saving endangered languages. His research interests include services computing and multiagent systems. He founded the Technical Committee on Services Computing in the Institute of Electronics, information and Communication Engineers (IEICE), in 2012. He received the Achievement Award of the Institute of Electronics, Information and Communication Engineers for this work, in 2013.

WINDA MONIKA received the bachelor's degree from Universitas Pendidikan Indonesia, in 2013, and the master's degree from the University of Tsukuba, Japan, in 2018. She is currently a Faculty Member with the Library Science Department, Faculty of Humanities, Universitas Lancang Kuning, Indonesia. Her current research interests include metadata, the semantic web, natural language processing, and digital humanities. She has contributed to various studies, including the development of metadata models for organizing digital archives and the application of latent semantic analysis for topic modeling in Indonesian children's literature.

ANGGI HANAFIAH received the bachelor's degree in computer science from the Institute of Information Management and Computer Science, in 2013, and the master's degree in computer science from Universitas Putra Indonesia, in 2014. He is currently a Lecturer with the Department of Informatics Engineering, Universitas Islam RIau, Indonesia. His current research interests include artificial intelligence, machine learning, and natural language processing.

Benchmarking Open-Source Large Language Models for Sentiment and Emotion Classification in Indonesian Tweets

ORIGINALITY REPORT

6%	5%	4%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	doaj.org Internet Source	3%
2	www.mdpi.com Internet Source	2%
3	Bhavana Verma, Priyanka Meel, Dinesh Kumar Vishwakarma. "Navigating sentiment analysis through fusion, learning, utterance, and attention Methods: An extensive four-fold perspective survey", Engineering Applications of Artificial Intelligence, 2025 Publication	1%

Exclude quotes	Off	Exclude matches	< 1%
Exclude bibliography	On		

Benchmarking Open-Source Large Language Models for Sentiment and Emotion Classification in Indonesian Tweets

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17