Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 3068~3078
e-ISSN: 2548-6861

Benchmarking Deepseek-LLM-7B-Chat and Qwen1.5-7B-Chat for Indonesian Product Review Emotion Classification

Galih Setiawan Nurohim[1]*, Heribertus Ary Setyadi[2]*, Ahmad Fauzi [3]**
* Information System, Universitas Bina Sarana Informatika
** Accounting Information System, Universitas Bina Sarana Informatika
galih.glt@bsi.ac.id [1], heribertus.hbs@bsi.ac.id [2], ahmad.fzx@bsi.ac.id [3]

Article Info

Article history:

Received 2025-09-27
Revised 2025-11-11
Accepted 2025-11-15

Keyword:

DeepSeek,
Emotion Classification,
LLM,
Qwen..

ABSTRACT

Upon completing their shopping experience on an e-commerce platform, users have the opportunity to leave a review. By analyzing reviews, businesses can gain insight into customer emotions, while researchers and policymakers can monitor social dynamics. Large Language Models (LLMs) utilization is identified as a promising methodology for emotion analysis. LLMs have revolutionized natural language processing capabilities, yet their performance in non-English languages, such as Indonesian, necessitates a comprehensive evaluation. This research objective is to perform a comprehensive analysis and comparison of Deepseek-LLM-7B-Chat and Qwen1.5-7B-Chat, two prominent open-source Large Language Models, for the emotion classification of Indonesian product reviews. By leveraging the PRDECT-ID dataset, this study evaluates the performance of both models in a few-shot learning scenario through prompt engineering. The methodology outlines the data preprocessing pipeline, a detailed few-shot prompt engineering strategy tailored to each model's characteristics, model inference execution, and performance assessment using the accuracy, precision, recall, and F1-score metrics. Analytical results reveal DeepSeek achieved an accuracy of 43.41%, exhibiting a considerably superior ability to comprehend instructions compared to Qwen, which attained a maximum accuracy of only 20.35% and often yielded near-random predictions. An in-depth error analysis indicates that this performance gap is likely attributable to factors such as pre-training data bias and tokenization mismatches with the Indonesian language. This research offers empirical evidence regarding the comparative strengths and weaknesses of DeepSeek and Qwen, providing a diagnostic benchmark that underscores the significance of instruction tuning and robust multilingual representation for Indonesian NLP tasks.

I. INTRODUCTION

Technological advancements have simplified shopping activities with various e-commerce platforms emergence that allow consumers to shop from home [1]. This also benefits sellers, enabling them to reach consumers more effectively through compelling content [2]. E-commerce platforms provide benefits like a quicker buying process, faster responses to customer or market needs, and a variety of payment methods [3]. Nonetheless, e-commerce platforms have their downsides, like customers not being able to see products in person and sellers not being able to understand what customers want in detail [4]. Customers on e-commerce sites can post reviews after they've finished shopping. Their feedback is often given as a star rating and a comment, which can contain rich emotional expressions [5].

Product reliability and available reviews can indicate a consumer's level of satisfaction with their purchased products [6]. Many user-generated reviews often include strong emotional content. This review data is highly promising and can be utilized by both customers and companies [7]. Customers can read reviews to learn more about a product's

quality. However, because there are so many reviews, it's hard to read all the customer feedback individually to get useful information [8]. Sellers need to look at the reviews to help their business grow. Addressing this issue requires a method to process and analyze data by representing it in a format comprehensible to computers [9]. An additional problem is fake reviews, which can cost e-commerce businesses money and trick customers into making a wrong decisions. Fake reviews also make it hard to know the real emotions in reviews from actual buyers. To solve this, we need to use text mining to analyze the review text given by e-commerce users [10].

For the task of emotion classification in text, which aims to identify specific emotional states such as joy, anger, or sadness, two primary methodological approaches are commonly employed. The first is a machine learning-based approach, which requires a labeled training dataset to automatically classify emotional expression [11]. Historically, sentiment analysis originated with classic machine learning approaches, such as Naive Bayes, which are straightforward but effective for short textual data. Krugmann (2024) research indicates that this method is still relevant for public opinion analysis on social media [12]. Nonetheless, its constraints in processing long contexts spurred the development of early deep learning models such as Long Short-Term Memory (LSTM), which demonstrated a superior ability to capture sequential dependencies in text and comprehend emotional subtleties [13]. Second, there is the lexicon-based approach, which requires a dictionary containing predefined lexicons and information on the polarity of sentiment-related words [14]. Following a study by Qorib, et al., who used the Textblob library for sentiment scoring, vectorized words with TF-IDF, and classified them using LinearSVC on a COVID-19 vaccine Twitter dataset, the proposed model achieved 96.7% accuracy [15].

The subsequent paradigm shift was defined by the introduction of the transformer architecture. Research by Ahmadian et al. (2023) and Jazuli et al. (2025) indicates that BERT and its variants can enhance the accuracy of Indonesian emotion classification, addressing both emotion classification and aspect-based analysis. Nonetheless, while BERT is effective, its methodology remains highly dependent on a pretraining + fine-tuning framework that requires a substantial amount of labeled data. Recent advancements are characterized by the advent of Large Language Models (LLMs), which are a huge step up from BERT [16] [17]. LLMs possess a much larger parameter capacity and excel in few-shot learning, enabling them to execute novel tasks like emotion classification with minimal or no extra training. Moreover, LLMs demonstrate emergent abilities that smaller models lack, such as more sophisticated reasoning and adapting to nuanced emotional contexts [18]. Consequently, LLMs are more versatile and flexible than earlier transformer models. In a recent 2025 research, Aydin et al. compared many popular LLMs like DeepSeek, Qwen, ChatGPT, Gemini, and Llama. They decided to focus on DeepSeek-LLM-7B-Chat and Qwen1.5-7B-Chat. The reason was that both models have a smaller, more manageable size (7B) for academic tests and have technical features that work well with non-Indo-European languages [19].

Developed by DeepSeek AI, DeepSeek is a powerful LLM built for handling multiple languages, specific applications, and tasks that need a lot of knowledge [20]. DeepSeek uses a very good training process and a large dataset to get better at factual accuracy, logical thinking, and being reliable in different NLP tests. Despite both models being based on the transformer architecture, their distinct training methodologies, optimization strategies, and application emphases require a comprehensive comparative analysis to assess their respective strengths, weaknesses, and appropriateness for a range of NLP tasks [21]. Nevertheless, LLM evaluations majority continue to focus on the English language. In fact, a model's performance really depends on the language it was trained on, and how well it can tokenize the target language [22]. Because Indonesian has a different word structure than Indo-European languages, it creates its own unique problems for NLP models.

Having datasets like PRDECT-ID makes this research even more important [23]. This dataset includes Indonesian product reviews with marked emotions, which is great for testing how well models work for emotion analysis in e-commerce [24]. Moreover, research that integrate LLMs capabilities with specific tasks, such as emotion classification, can provide dual contributions: advancing NLP knowledge for non-English languages, and directly addressing industry requirements like market research, marketing strategy development, and enhancing customer experience [25]. With this in mind, this research aims to fill a research gap by comparing two popular open-source LLMs, DeepSeek-LLM-7B-Chat and Qwen1.5-7B-Chat. The main goal is to see how well they perform at classifying emotions in Indonesian product reviews, using the PRDECT-ID dataset. Although large-scale language models (LLMs) have achieved state-of-the-art performance in English and Chinese, their effectiveness in processing Indonesian text remains limited due to the scarcity of linguistic resources and insufficient domain adaptation. This gap has practical implications: the low accuracy of LLMs in understanding Indonesian text can reduce the reliability of automated opinion analysis systems, introduce bias in business decision-making, and hinder the deployment of AI-based public service applications. This observation aligns with the findings of Tohir et al. (2024), who emphasized that Indonesian-language NLP still faces significant challenges in large-scale model comprehension and retrieval-augmented learning contexts. [26] . Previous benchmarking efforts, such as the BHASA study[27], revealed that state-of-the-art large language models like GPT-3.5 and GPT-4 still underperform in Southeast Asian languages, including Indonesian, particularly in linguistic comprehension and cultural sensitivity. However, their evaluation was limited to closed-weight models and did not include newer open-weight multilingual models such as

DeepSeek and Qwen. This limitation highlights the need for further benchmarking of open-access LLMs using Indonesian datasets to better assess their capability in local emotion analysis contexts.

The selection of DeepSeek-LLM-7B-Chat and Qwen1.5-7B-Chat in this study was motivated by their multilingual capacity, open-weight accessibility, and inference efficiency, which make them suitable for academic benchmarking and resource-constrained deployment environments. The selection was further informed by practical deployment considerations: both models support quantization and efficient inference techniques that enable execution on mid-range GPUs commonly available in academic research laboratories, thereby improving accessibility for reproducible research and resource-limited environments [28] [29]. The objective of this research is to diagnostically compare the effectiveness of the DeepSeek-LLM-7B-Chat and Qwen1.5-7B-Chat models for emotion classification, based on the F1-Score, precision, and recall metrics. While the study's contribution is primarily empirical, it provides a foundational benchmark for subsequent methodological research, particularly in optimizing prompt engineering strategies, understanding cross-lingual model biases, and exploring quantization-based deployment feasibility under constrained computing environments such as Google Colab free-tier GPUs. The insights gained also help explain potential causes behind low model accuracy, including tokenization mismatch, prompt ambiguity, and linguistic data scarcity, thereby positioning this study as a reproducible reference for future Indonesian LLM evaluation. Critically, this study does not merely present performance scores but seeks to interpret the underlying reasons for success and failure, offering insights into the challenges of applying general-purpose LLMs to a specific, low-resource language context like Indonesian.

II. METHOD

This study was initiated by collecting data from the PRDECT-ID dataset, comprising 5,400 Indonesian e-commerce product reviews annotated with five key emotion labels: Anger, Fear, Happy, Love, and Sadness. It culminated in an analysis of the results to compare the performance between the models. Data collection for this study was performed using the PRDECT-ID dataset. The dataset comprises 5,400 e-commerce product reviews in the Indonesian language, each annotated with one of five core emotion labels: Anger, Fear, Happy, Love, and Sadness. The subsequent step involved data pre-processing, encompassing simple text cleaning procedures. These included normalizing all characters to lowercase to ensure consistency and tokenization to prepare the textual data for model input. The next stage was model selection, where two large language models, DeepSeek-LLM-7B-Chat and Qwen1.5-7B-Chat, were chosen for this study. Figure 1 illustrates this research stages.

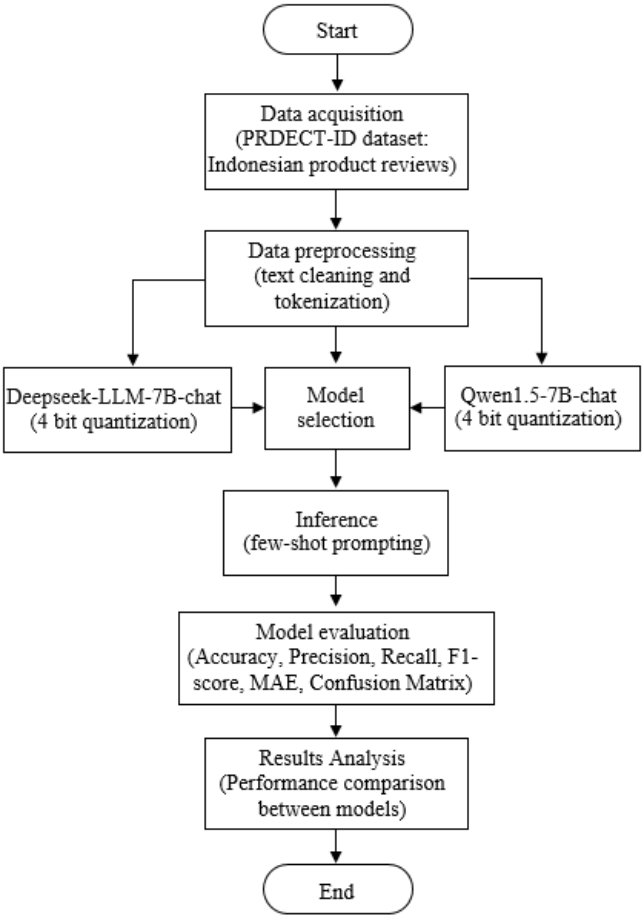

Figure 1 Research stages.

To optimize GPU memory usage, both models were loaded using a 4-bit quantization technique with a BitsAndBytes configuration. The inference was carried out on a Google Colab instance, leveraging a Tesla T4 GPU to strike a balance between computational capacity and resource constraints. For performance assessment, this research employed a number of evaluation metrics, including Accuracy, Balanced Accuracy, Precision, Recall, F1-score, and Mean Absolute Error (MAE). In addition, a Confusion Matrix was utilized to illustrate the distribution of predictions across each class. The concluding step involved a results analysis, assessing the performance of two models across both accuracy and inference speed. The analysis also encompassed an interpretation of performance for each emotion class, which allowed for an examination of the trade-off between predictive quality and computational efficiency.

A. Dataset

For this research, the PRDECT-ID dataset, developed by Sutoyo et al., was employed. The dataset comprises 5,400 Indonesian product reviews that were gathered from the Tokopedia platform. Stored in a .csv format, the data is publicly accessible through Mendeley Data for academic use. Each entry within the dataset provides the customer review text, a sentiment label (positive or negative), and an emotion

label (Anger, Fear, Happy, Love, Sadness). The dataset further encompasses additional attributes pertaining to the product, seller, and customer rating. In total, the dataset comprises 5,400 annotated samples distributed across five emotion categories: Love (809; 15.0%), Happiness (1,770; 32.8%), Anger (699; 13.0%), Fear (920; 17.0%), and Sadness (1,202; 22.3%). This proportional distribution demonstrates a reasonable balance of emotional representation across Indonesian product reviews, thereby supporting robust model evaluation and comparison. Table 1 presents the emotion annotation criteria, and Table 2 describes the list of dataset attributes. The distribution of emotions in the dataset is depicted in Figure 1, and Figure 2 shows the distribution of emotion labels by product category.

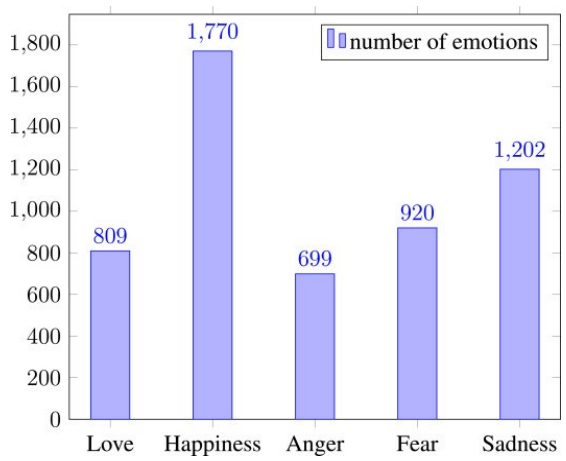

Figure 2. Distribution of emotion labels in the PRDECT-ID dataset [23]

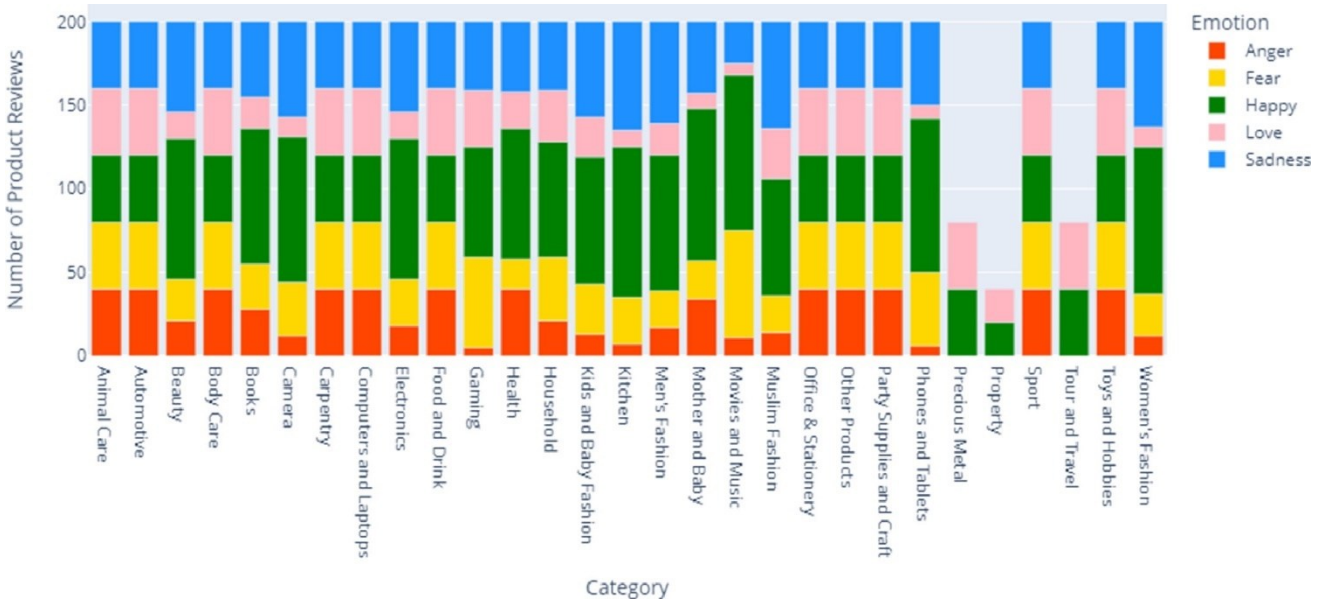

Figure 3. Distribution of emotion labels based on product category (Sutoyo et al)

B. Preprocessing

The PRDECT-ID dataset was subjected to the following basic preprocessing steps prior to model training:

1) *Column Selection*

For the purpose of this analysis, only the 'Customer Review' and 'Emotion' columns were selected from the dataset, with all other columns being excluded.

2) *Label Normalization*

For consistency, the emotion labels were normalized to a title-case format (e.g., 'happy' → 'Happy')

3) *Emotion Label Determination*

From the dataset, five emotion classes were identified: Anger, Fear, Happy, Love, and Sadness.

C. Testing Scenarios and Prompt Design

This research utilized distinct prompt designs for the DeepSeek and Qwen models to suit their respective decoding characteristics.

1) *Deepseek (Few-shot Prompt Configuration)*

- Prompt format: A short definition of each emotion followed by explicit examples (deterministic few-shot).
- Output Instruction: he model was instructed to provide only a single-word label corresponding to one of the five emotions.
- Output Instructions: are asked to provide only one-word label answers without additional text.
- Parsing Output: Conducted using a regex pattern to match emotion labels. If no valid label was detected, the fallback class was set to “Happy.”

- Max token length: limited to 256 tokens to ensure deterministic decoding and to avoid truncation in Google Colab GPU memory.
- Prompt Example:

Task: Identify the emotion expressed in the text below.
Possible emotions: Anger, Fear, Happy, Love, Sadness.

Example 1:
Text: "Produk ini bagus sekali, saya sangat puas."
Label: Happy

Example 2:
Text: "Saya kecewa, barangnya rusak saat diterima."
Label: Anger

Text: "Pelayanan toko ini lambat dan tidak ramah."
Label: .

2) *Qwen (Experiment 1 – Few-shot Deterministic)*

- Prompt Format: Fixed example order for all five emotions to ensure stable inference results.
- Instruction: The model was explicitly told to output one emotion label word only, matching one of {Anger, Fear, Happy, Love, Sadness}.
- Parsing output: Similar to DeepSeek, a regex-based extraction ensured valid label output; invalid responses were replaced with the nearest valid emotion based on semantic similarity.
- Max token length: 512 tokens, providing room for multi-example context while remaining memory-efficient on a 24 GB GPU.
- Prompt Example:

Anda adalah sistem klasifikasi emosi untuk ulasan produk e-commerce berbahasa Indonesia.
Label valid: Anger, Fear, Happy, Love, Sadness.
Pilih hanya SATU label.

Contoh (acak setiap kali):
T: "Sebaiknya rekam unboxing, takutnya ada yang kurang."
Label: Fear

T: "Produknya bagusss, suka bangett!!!"
Label: Love

T: "Sangat kecewa, tidak lengkap."
Label:

3) *Qwen (Experiment 2 – Few-shot Randomized)*

- Prompt Format: Same content as the deterministic variant, but the order of examples was randomized at each inference call to test output stability and bias sensitivity.
- Randomization: Implemented using Python's random.shuffle() to permute five few-shot examples per input.
- Output instruction: Single-word emotion label only, identical to deterministic mode.
- Parsing output: Regex validation applied; invalid or multi-word outputs defaulted to the most frequent emotion in the dataset.
- Max token length: 512 tokens for balanced memory and linguistic coverage
- Prompt Example:

D. Model Quantization

As the substantial model size (7B parameters) exceeded the memory capacity of the Tesla T4 GPU, a 4-bit quantization technique using the BitsAndBytes library was utilized. The configuration applied is:

- load_in_4bit=True to load the model weights in 4-bit format.
- bnb_4bit_use_double_quant=True for more efficient quantization.
- bnb_4bit_quant_type="nf4" for the normalization quantization scheme.
- bnb_4bit_compute_dtype=torch.float16 for lighter yet stable computation.

To optimally distribute the computational load between the GPU and CPU, the device_map="auto" and max_memory={0: "6GiB", "cpu": "30GiB"} parameters were implemented. This approach enabled the large DeepSeek-LLM-7B-Chat model to execute on a T4 GPU in Google Colab without causing an out-of-memory error.

E. Experimental Configuration

As All experiments were conducted in a Google Colab free-tier environment, which provides a single NVIDIA Tesla T4 GPU (16 GB VRAM), 12 GB of system RAM, and limited persistent storage.

To ensure that both large-scale models (DeepSeek-LLM-7B-Chat and Qwen1.5-7B-Chat) could be executed under these constraints, several configurations were applied as follows:

- Environment Setup
  Platform: Google Colab Free
  GPU: NVIDIA Tesla T4 (16 GB VRAM)
  CUDA version: 12.4
  Python version: 3.10
  PyTorch version: 2.3.0 + cu121
  Transformers version: 4.42.0
  BitsAndBytes version: 0.43.1
- Hardware Management
  The device_map="auto" parameter was employed to automatically distribute model layers between the GPU and CPU.
  A custom configuration of max_memory={0: "6GiB", "cpu": "30GiB"} was defined to prevent GPU memory overflow.
  This strategy was critical for ensuring stable inference performance, especially for the 7B-parameter DeepSeek-LLM model.
- Model Loading and Quantization Consistency

Both models were loaded using identical 4-bit quantization settings to maintain fairness and efficiency across experiments:
load_in_4bit=True
bnb_4bit_use_double_quant=True
bnb_4bit_quant_type="nf4"
bnb_4bit_compute_dtype=torch.float16
This configuration allowed the models to fit within Colab's limited GPU memory without compromising output consistency.

- Prompt Configuration
  Each model was evaluated using a few-shot prompting strategy for emotion classification, with a maximum token length of 256 to prevent truncation.
  For DeepSeek, a deterministic few-shot prompt was used.
  For Qwen, two prompting variations were tested:

Experiment 1: deterministic few-shot without randomization.
Experiment 2: randomized few-shot order using random.shuffle() to evaluate stability and generalization.

- Inference Parameters
  The inference configuration was standardized across all models to ensure reproducibility:
  do_sample=False
  temperature=0.0
  top_p=1.0
  max_new_tokens=10
  pad_token_id=tokenizer.eos_token_id
  The fallback rule assigned the majority label (*Happy*) if no valid emotion label was detected through regex parsing.

E. Evaluation Metrics

Model performance was evaluated using standard classification metrics: accuracy, balanced accuracy (to handle imbalanced classes), precision, recall, and F1-score.

III. Result and Discussion

A. Quantitative Performance Results

The experiment results indicate a significant difference in performance between Deepseek-LLM-7B-Chat and Qwen1.5-7B-Chat (Table I).

TABLE I
Overall Performance Comparison: DeepSeek vs. Qwen

Model	Accuracy	Balanced Accuracy	Speed (it/s)	Total time (5400 data)
DeepSeek	43.41%	38.93%	1.60	~56 menit
Qwen (Experiment 1)	17.04%	20.00%	2.34	~38 menit
Qwen (Experiment 2)	20.35%	20.32%	1.98	~45 menit

As shown in Table I, the experiment results reveal that Deepseek-LLM-7B-Chat outperformed Qwen1.5-7B-Chat in Indonesian emotion classification. DeepSeek achieved an overall accuracy of 43.41%, almost double Qwen's accuracy in the first (17.04%) and second (20.35%) trials. This suggests that DeepSeek has a better ability to understand few-shot prompt-based classification instructions for product reviews. From the perspective of balanced accuracy, the performance difference becomes even more apparent. DeepSeek achieved 38.93%, while Qwen only reached around 20% despite using two prompt variations. Qwen's lower balanced accuracy indicates that the model tends to be biased towards majority classes (e.g., Happy) and is less capable of detecting minority classes such as Fear or Love.

Regarding computational efficiency, Qwen exhibits faster inference speed compared to DeepSeek. The initial experiment with Qwen only took approximately 38 minutes for 5,400 data points, while DeepSeek required 56 minutes. This is understandable, given that DeepSeek processes a more complex computational workload with extensive prompt processing and stricter decoding.

TABLE II
DeepSeek Model Classification Report

Class	Precision	Recall	F1-Score	Support
Anger	0.306	0.335	0.320	699
Fear	0.335	0.289	0.310	920
Happy	0.681	0.538	0.601	1770
Love	0.658	0.131	0.219	809
Sadness	0.344	0.653	0.451	1202
Macro Avg	0.465	0.389	0.380	5400

Table II, which provides the detailed classification report for DeepSeek, reveals that the model has strengths in the Happy class (F1-score 0.601) and the Sadness class (Recall 0.653). This indicates that DeepSeek is reasonably effective in recognizing reviews with sentiments of satisfaction or disappointment. However, a weakness is evident in the 'Love' class, with a very low recall (0.131). The model frequently miscategorizes 'Love' as 'Happy,' indicating a difficulty in distinguishing between general expressions of satisfaction and overly enthusiastic expressions.

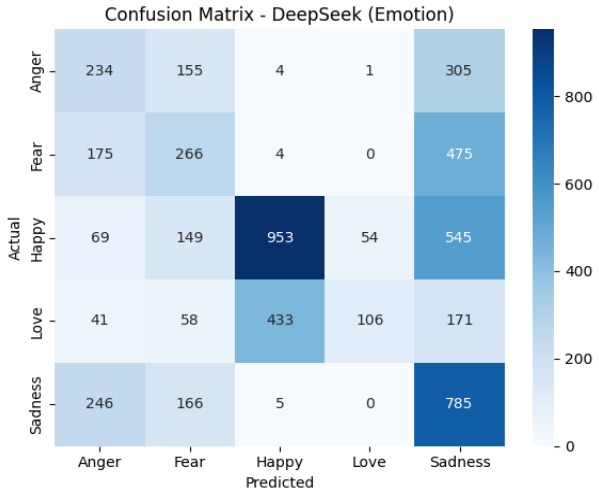

Figure 4. Confusion Matrix Deepseek

TABLE III
MODEL QWEN REPORT

Class	Precision	Recall	F1-Score	Support
Anger	0.122	0.187	0.147	699
Fear	0.170	0.196	0.182	920
Happy	0.320	0.199	0.246	1770
Love	0.165	0.219	0.188	809
Sadness	0.238	0.215	0.226	1202
Macro Avg	0.203	0.203	0.198	5400

Qwen's classification report in Table III shows a much lower performance across all classes. The highest F1-score was only achieved in the 'Happy' class (0.246), while other classes were below 0.2. With a macro average F1 of only 0.198, Qwen was far behind DeepSeek (0.380). This further confirms that Qwen, in its tested configuration, is not yet optimal for Indonesian emotion classification tasks that employ a few-shot prompting strategy.

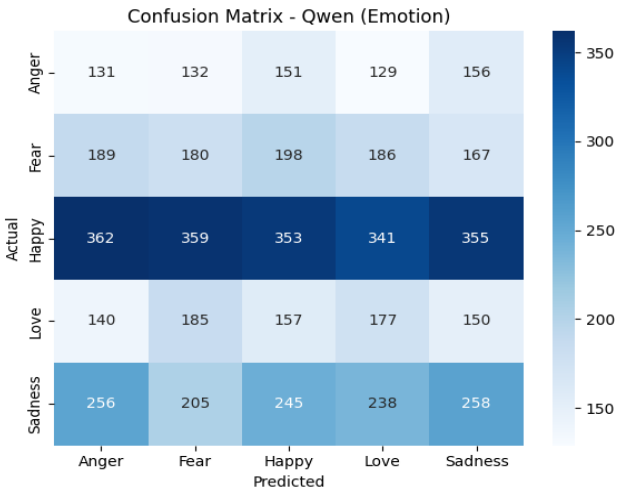

Figure 5. Confusion Matrix QWEN

Test results reveal a significant disparity between DeepSeek-LLM-7B-Chat and Qwen1.5-7B-Chat in the emotion classification task on the PRDECT-ID dataset.
1) Global Performance: DeepSeek Is More Reliable.

B. *Statistical Significance of Performance Difference*

To empirically validate that the observed performance difference between DeepSeek and Qwen is not due to random chance, we conducted McNemar's test. This test is suitable for paired nominal data, as in this case where both models classified the same 5,400 product reviews. The test specifically focuses on the discordant cases where the models disagree to determine if their success proportions are significantly different. The following contingency table summarizes the predictions from both models.

TABLE IV
CONTINGENCY TABLE OF MODEL PREDICTIONS

	Qwen1.5-7B-Chat Correct	Qwen1.5-7B-Chat Incorrect	Total
DeepSeek-LLM-7B-Chat Correct	475	**1869**	2344
DeepSeek-LLM-7B-Chat Incorrect	**624**	2432	3056
Total	1099	4301	5400

The results of the McNemar's test, which focuses on these two numbers, are as follows:Chi-Square (χ^2) Statistic: 620.75

- Chi-Square Chi-Square (χ^2) Statistic: 620.75
- P-value : 5.12×10^{-137}

With a p-value far below the significance threshold of 0.05, we reject the null hypothesis that both models have the same error rate. The test provides overwhelming evidence that the performance difference is real. Specifically, DeepSeek was correct in 1,869 instances where Qwen failed, whereas Qwen was correct in only 624 instances where DeepSeek failed. Therefore, we confidently conclude that DeepSeek-LLM-7B-Chat is statistically superior to Qwen1.5-7B-Chat for this task.

C. *Analysis and Discussion*

Test results reveal a significant disparity between DeepSeek-LLM-7B-Chat and Qwen1.5-7B-Chat in the emotion classification task on the PRDECT-ID dataset.

1) Interpretation of Sub-50% Accuracy and Practical Viability.

The experimental results unequivocally reveal a significant disparity between Deepseek-LLM-7B-Chat and Qwen1.5-7B-Chat in the emotion classification task on the PRDECT-ID dataset. Overall, DeepSeek achieves an accuracy of 43.41%, a figure more than double Qwen's best performance of 20.35%. Qwen's achievement is particularly telling as it nearly mirrors the random guessing baseline for a five-class problem (1/5 = 20%). This strongly suggests that the model was fundamentally unsuccessful in utilizing the few-shot instruction context provided in Indonesian, operating almost as a stochastic parrot rather than a comprehending agent. In contrast, DeepSeek, despite its limitations, managed to extract more meaningful patterns from the prompts and data, indicating a superior capability for cross-lingual instruction following. However, it is crucial to contextualize this "superiority"; while statistically significant, an accuracy of 43.41% remains far below the performance level of a specialized, fine-tuned model like IndoBERT, which typically achieves scores above 80% on similar tasks. This immediately raises questions of practical viability. In a real-world e-commerce system, deploying a model with over 50% error rate could lead to incorrect customer routing, flawed market analysis, and ultimately, business losses. Therefore,

DeepSeek's performance should be regarded not as a solution, but as a critical diagnostic baseline highlighting the immense gap between general-purpose LLMs and task-specific, language-adapted models. A McNemar's test confirmed this performance gap is statistically significant ($p < 0.001$), providing robust empirical validation that DeepSeek's superiority is not an artifact of random chance but a reproducible phenomenon rooted in the models' architectures and training

2) Per-Class Performance and Model-Specific Behavioral Patterns

Moving beyond aggregate metrics, an analysis of performance per class reveals distinct and problematic behavioral patterns for each model, which can be directly linked to their underlying linguistic and cognitive processing capabilities.

DeepSeek-LLM-7B-Chat's Strengths and Systematic Biases: The confusion matrix for DeepSeek (Figure 4) paints a picture of a model that understands broad sentiment but fails at granular emotion detection. It shows a strong inclination to identify the 'Happy' class (True Positives = 953), its strongest performing category. However, this strength is coupled with a significant and systematic bias: it frequently miscategorizes 'Love' as 'Happy' (433 cases). This error pattern is not random; it suggests the model struggles to differentiate between general satisfaction (e.g., "barangnya bagus") and more intense, affectionate expressions (e.g., "saya jatuh cinta sama produk ini"). The model's semantic space for positive sentiment appears to be compressed, lumping together distinct emotional states. Furthermore, for the 'Sadness' class, while recall is relatively high (785), the predictions are often conflated with 'Anger' (246) and 'Fear' (166). This suggests indistinct boundaries among negative emotions. In Indonesian, expressions of disappointment ("kecewa"), anger ("marah"), or fear ("takut") can share lexical items or appear in similar contexts (e.g., a product not arriving), and DeepSeek appears unable to disambiguate these subtle cues.

Qwen1.5-7B-Chat's Failure to Internalize the Classification Task: Qwen's behavior (Figure 5) is even more dysfunctional and indicative of a deeper failure. Its confusion matrix exhibits a 'spread out' and uniform distribution of predictions across all classes. For example, for the true 'Fear' class, its predictions were dispersed almost evenly among all five labels (around 180–198 per category). This phenomenon is the hallmark of a model that has not learned the decision boundaries of the classification task. It elucidates why Qwen's balanced accuracy remained stagnant at 20%—the model was unsuccessful in internalizing the classification instructions and tended to provide generic, near-random responses. This pattern, consistent across different prompt variations, suggests a fundamental limitation in the model's ability to comprehend and execute tasks in Indonesian, a hypothesis that is empirically tested and confirmed in the subsequent section

3) Empirical Justification: Linguistic Mismatch as the Root Cause

TABLE V
TOKENIZATION AND VOCABULARY COVERAGE ANALYSIS

Metric	Qwen1.5-7B-Chat	DeepSeek-LLM-7B-Chat
Average Raw Tokens	32.97	39.33
Average Prompt Tokens	205.71	250.99
Top-1000 Word Coverage (%)	23.20%	20.90%
T-test (Prompt Token Length Difference)	$t = 77.808$, $p < 0.001$	

The most striking finding is that DeepSeek generated significantly longer prompt representations (avg. 250.99 tokens) compared to Qwen (avg. 205.71 tokens), a difference that is highly statistically significant ($p < 0.001$). This is not a trivial detail. It suggests that DeepSeek's tokenizer is better suited for Indonesian, preserving more semantic and morphological detail from the input text. In contrast, Qwen's tokenizer, likely optimized for Chinese and English, compresses Indonesian text more aggressively. This aggressive compression can lead to the fragmentation of multi-word expressions or the discarding of crucial affixal information (like prefix me-, ber-, suffix -kan, -an) that are fundamental to Indonesian grammar and meaning. For example, a word like "kekecewaanku" (my disappointment) might be tokenized into meaningful sub-words by DeepSeek but treated as a single, unknown entity by Qwen.

Furthermore, the vocabulary coverage analysis is damning. Qwen's tokenizer covered only 23.2% of the top-1000 most frequent words in the dataset. This limited coverage highlights a profound domain and language mismatch between the models' pretraining corpora (predominantly Chinese and English) and the Indonesian review dataset. The model's near-random accuracy and uniform confusion matrix are now no longer a mystery; they are a direct symptom of this linguistic disconnect. The model simply fails to "see" or "understand" the key words that signal specific emotions. In contrast, DeepSeek, while still imperfect, demonstrated stronger adaptability, likely due to more extensive multilingual exposure during its instruction tuning phase, allowing it to better generalize to Indonesian.

.

4) Qualitative Error Analysis: Concrete Examples of Semantic and Pragmatic Failure

To complement the quantitative analysis and provide a deeper understanding of the models' limitations, a manual inspection of misclassified examples was conducted. This qualitative analysis illuminates the semantic and pragmatic shortcomings of both DeepSeek and Qwen, revealing specific failure patterns that aggregate metrics obscure. The analysis reveals several distinct patterns, which are categorized and presented below.

A. Systematic Misclassification of Genuinely Positive Reviews

A striking and consistent failure mode, particularly for DeepSeek, is the systematic misclassification of genuinely positive reviews as negative emotions. As shown in Table VI.A, reviews containing explicit positive sentiment were frequently labeled as 'Sadness' or 'Fear'.

TABLE VI
EXAMPLES OF POSITIVE REVIEWS MISCLASSIFIED AS NEGATIVE

Review Text	True Label	DeepSeek Prediction	Qwen Prediction
"Alhamdulillah berfungsi dengan baik. Packaging aman. Respon cepat dan ramah. Seller dan kurir amanah"	Happy	Sadness	Fear
"bagus. berjalan baik dan cukup silent. kokoh dan muat untuk laptop 16 inch"	Happy	Sadness	Sadness
"Barang cepat sampai. Kualitas bagus dan pengiriman aman tidak ada yang pecah. Manstapu ??????"	Happy	Sadness	Fear

Analysis: This error pattern is particularly concerning as it represents a fundamental inversion of sentiment. The models appear to be unable to correctly process positive Indonesian expressions, possibly due to a severe bias in their pre-training data or a fundamental tokenization error where positive cultural or religious expressions like "Alhamdulillah" are misinterpreted. The failure to recognize simple positive descriptors like "bagus" and "berjalan baik" points to a profound disconnect between the models' learned representations and the Indonesian language's lexical semantics for satisfaction.

B. The 'Love' vs. 'Happy' Conflation

A more subtle, yet significant, error pattern is the conflation of the 'Love' emotion with the more general 'Happy'. 'Love' in product reviews implies a stronger, more affectionate endorsement, whereas 'Happy' denotes general satisfaction. As shown in Table VII, both models consistently fail to capture this intensity difference.

TABLE VII
EXAMPLES OF 'LOVE' MISCLASSIFIED AS 'HAPPY'

Review Text	True Label	DeepSeek Prediction	Qwen Prediction
"recommended, fast print fast response"	Love	Happy	Happy
"Recommended Seller!! Respon cepat, pengiriman cukup cepat..."	Love	Happy	Happy

Analysis: The models fail to recognize that strong endorsements like "recommended" (repeated for emphasis) and "Recommended Seller!!" signal an emotion beyond simple satisfaction. This indicates a lack of semantic granularity in the models' understanding of positive sentiment. They operate on a binary positive/negative axis rather than a nuanced emotional spectrum, lumping distinct positive states into a single, dominant 'Happy' class.

C. Inability to Distinguish Negative Emotions

The models also demonstrate a limited ability to differentiate between distinct negative emotions, particularly 'Anger' and 'Sadness'. 'Anger' often involves blame, frustration, and a sense of being wronged, while 'Sadness' is more passive, linked to disappointment. Table VIII shows that the models often misinterpret expressions of clear frustration as mere disappointment.

TABLE VIII
EXAMPLES OF 'ANGER' MISCLASSIFIED

Review Text	True Label	DeepSeek Prediction	Qwen Prediction
"Sangat mengecewakan, barang yg dikirim tidak sesuai konfirmasi, sampai harus bolak-balik 2 kali...	Anger	Sadness	Happy
"Recommended Seller!! Respon cepat, pengiriman cukup cepat..."	Anger	Sadness	Sadness

Analysis: The misclassification of a clearly angry review, which includes strong negative markers like "Sangat Mengecewakan" and specific complaints about service, as 'Sadness' or even 'Happy' indicates that the models are not capturing key discourse markers of frustration and blame. This failure to differentiate between active frustration and passive disappointment limits their utility for applications that need to identify and urgently address customer anger.

D. Gross Misclassification of Short, Context-Dependent Reviews

The most severe errors involve the complete inversion of sentiment, often with very short reviews where context is paramount. Table IX presents a critical example of this failure mode.

TABLE IX
EXAMPLE OF GROSS SENTIMENT INVERSION

Review Text	True Label	DeepSeek Prediction	Qwen Prediction
"Sekali pakai"	Fear	Happy	Happy

Analysis: The review "Sekali pakai" (Used once), labeled as 'Fear' (implying a fear that the product will break after a single use), was predicted as 'Happy' by both models. This catastrophic failure suggests the models are overly reliant on superficial cues or are completely baffled by short, context-dependent phrases. Without explicit emotional keywords, they default to a positive prediction, highlighting a dangerous

unreliability when processing concise or implicit feedback. This qualitative evidence confirms that without significant adaptation, such as targeted fine-tuning or improved tokenization, general-purpose LLMs remain unreliable for the nuanced task of emotion classification in Indonesian.

5) Limitations of Generative LLMs and Avenues for Future Research

These findings substantiate an inherent limitation associated with employing general-purpose, generative LLMs for discriminative classification without subsequent, task-specific adaptation. The 43.41% accuracy achieved by Deepseek-LLM-7B-Chat is a robust baseline for few-shot scenarios, but it also serves as a clear indicator of the ceiling for this approach. This study's primary contribution is not a novel method, but a clear, empirical benchmark that diagnoses specific failure modes, providing a valuable reference for future research. Based on our findings, we propose several promising avenues:

Indonesian-Centric Tokenization: Future work must prioritize the development or fine-tuning of tokenizers specifically on large Indonesian corpora. Techniques like SentencePiece or Byte-Pair Encoding (BPE) trained on Indonesian text would likely alleviate the fragmentation issue and improve vocabulary coverage, providing a stronger foundation for any model.

Domain-Specific Fine-Tuning: The most direct path to higher accuracy is to move beyond few-shot prompting and fine-tune these models on the PRDECT-ID dataset or a larger corpus of Indonesian reviews. Fine-tuning forces the model to learn the specific mapping between the Indonesian emotional lexicon and the target labels, a process far more effective than relying on the model's latent, cross-lingual knowledge.

Retrieval-Augmented Generation (RAG): A RAG framework could be highly effective. When the model encounters an ambiguous phrase like "takutnya nyesel," a RAG system could retrieve a database of similar phrases and their correct labels, effectively providing a "hint" during inference and grounding the model's prediction in relevant examples.

Cultural and Pragmatic Nuance Modeling: Ultimately, achieving high performance will require models that understand not just words, but cultural context. Indonesian communication often involves indirectness, politeness, and pragmatics that are not captured by surface-level text analysis. Future research should explore how to train or prompt models to recognize these subtle cues, which are critical for accurate emotion detection.

IV. CONCLUSION

This research performed a comparative analysis between the LLM models Deepseek-LLM-7B-Chat and Qwen1.5-7B-Chat for the task of emotion classification on product reviews written in Indonesian. The study's findings indicate that Deepseek-LLM-7B-Chat is significantly more effective, with an accuracy of 43.41%, whereas Qwen1.5-7B-Chat only managed to reach a maximum accuracy of 20.35%. The superiority of Deepseek-LLM-7B-Chat stems from its enhanced ability to follow complex few-shot instructions, which is likely a result of richer multilingual pre-training data and more effective instruction tuning. Conversely, Qwen1.5-7B-Chat exhibited performance that neared random chance, suggesting its incompatibility with Indonesian language tasks in a zero/few-shot setting, likely due to linguistic bias and tokenization issues.

The analysis revealed that even the superior model, DeepSeek, struggles with nuanced emotional distinctions, particularly confusing 'Love' with 'Happy', and performs poorly on minority classes. The consistently low accuracy across both models highlights a critical gap: general-purpose LLMs, without specific adaptation, are not yet practically viable for Indonesian sentiment classification. This study's primary contribution is not a new method, but a clear, empirical benchmark that diagnoses these specific failure modes, providing a valuable reference for future research into prompt optimization, fine-tuning, or retrieval-augmented approaches tailored for the Indonesian language. Although it is slower, the substantially higher accuracy of Deepseek-LLM-7B-Chat positions it as the more viable baseline for such future investigations.

ACKNOWLEDGEMENT

The authors gratefully acknowledge Bina Sarana Informatika University for financial support through the Bipemas grant program.

REFERENCES

[1] Khoirotulmuadiba Purifyregalia, Khothibul Umam, Nur Cahyo Hendro Wibowo, and Maya Rini Handayani, "Detecting Fake Reviews in E-Commerce: A Case Study on Shopee Using Support Vector Machine and Random Forest," *J. Appl. Informatics Comput.*, vol. 9, no. 3, pp. 955–965, 2025, doi: 10.30871/jaic.v9i3.9514.

[2] A. Daza, N. D. González Rueda, M. S. Aguilar Sánchez, W. F. Robles Espíritu, and M. E. Chauca Quiñones, "Sentiment Analysis on E-Commerce Product Reviews Using Machine Learning and Deep Learning Algorithms: A Bibliometric Analysisand Systematic Literature Review, Challenges and Future Works," *Int. J. Inf. Manag. Data Insights*, vol. 4, no. 2, 2024, doi: 10.1016/j.jjimei.2024.100267.

[3] M. R. R. Rana, A. Nawaz, T. Ali, A. M. El-Sherbeeny, and W. Ali, "A BiLSTM-CF and BiGRU-based Deep Sentiment Analysis Model to Explore Customer Reviews for Effective Recommendations," *Eng. Technol. Appl. Sci. Res.*, vol. 13, no. 5, pp. 11739–11746, 2023, doi: 10.48084/etasr.6278.

[4] P. S. Ghatora, S. E. Hosseini, S. Pervez, M. J. Iqbal, and N. Shaukat, "Sentiment Analysis of Product Reviews Using Machine Learning and Pre-Trained LLM," *Big Data Cogn. Comput.*, vol. 8, no. 12, 2024, doi: 10.3390/bdcc8120199.

[5] Kirtika, "Intelligent Systems And Applications In Enhancing Sentiment Classification Accuracy of Amazon Product Reviews via NLP Approaches," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 4, pp. 5752–5760, 2024, [Online]. Available: https://ijisae.org/index.php/IJISAE/article/view/7601

[6] N. Nabila and C. I. Ratnasari, "Topic Modeling of Skincare Comments from Female Daily," *J. Appl. Informatics Comput.*, vol. 9, no. 4, pp. 1394–1405, 2025, doi: 10.30871/jaic.v9i4.9625.

[7] P. Paul, S. Acharya, B. Misra, S. Majumder, N. Dey, and P. Pise, "Sentiment Analysis for E-Commerce Product Reviews Using CNN-LSTM," *2024 1st Int. Conf. Women Comput. InCoWoCo 2024 - Proc.*, no. November 2024, pp. 14–15, 2024, doi: 10.1109/InCoWoCo64194.2024.10863425.

[8] M. A. Kausar, S. O. Fageeri, and A. Soosaimanickam, "Sentiment Classification based on Machine Learning Approaches in Amazon Product Reviews," *Eng. Technol. Appl. Sci. Res.*, vol. 13, no. 3, pp. 10849–10855, 2023, doi: 10.48084/etasr.5854.

[9] A. Godia and L. K. Tiwari, "Sentiment Analysis and Classification of Product Reviews: A Comprehensive Study Using NLP and Machine Learning Techniques," *10th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2024*, vol. 1, no. March 2024, pp. 1247–1252, 2024, doi: 10.1109/ICACCS60874.2024.10717296.

[10] O. Shobayo, S. Sasikumar, S. Makkar, and O. Okoyeigbo, "Customer Sentiments in Product Reviews: A Comparative Study with GooglePaLM," *Analytics*, vol. 3, no. 2, pp. 241–254, 2024, doi: 10.3390/analytics3020014.

[11] K. A. F. A. Samah, N. F. A. Misdan, M. N. H. H. Jono, and L. S. Riza, "The Best Malaysian Airline Companies Visualization through Bilingual Twitter Sentiment Analysis: A Machine Learning Classification," *Int. J. Informatics Vis.*, vol. 6, no. 1, pp. 130–137, 2022, doi: 10.30630/joiv.6.1.879.

[12] J. O. Krugmann and J. Hartmann, "Sentiment Analysis in the Age of Generative AI," *Cust. Needs Solut.*, vol. 11, no. 1, 2024, doi: 10.1007/s40547-024-00143-4.

[13] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 4, p. 102048, 2024, doi: 10.1016/j.jksuci.2024.102048.

[14] P. Kumar and M. Kumar, "Review and Analysis of Product Review Sentiment Analysis using Improved Machine Learning Techniques," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 11, no. 10, pp. 946–951, 2023.

[15] M. Qorib, T. Oladunni, M. Denis, E. Ososanya, and P. Cotae, "Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset," *Expert Syst. Appl.*, vol. 212, no. January 2022, p. 118715, 2023, doi: 10.1016/j.eswa.2022.118715.

[16] A. Jazuli, Widowati, and R. Kusumaningrum, "Optimizing Aspect-Based Sentiment Analysis Using BERT for Comprehensive Analysis of Indonesian Student Feedback," *Appl. Sci.*, vol. 15, no. 1, pp. 1–28, 2025, doi: 10.3390/app15010172.

[17] H. Ahmadian, T. F. Abidin, H. Riza, and K. Muchtar, "Transformer-Based Indonesian Language Model for Emotion Classification and Sentiment Analysis," *Proceeding - Int. Conf. Inf. Technol. Comput. 2023, ICITCOM 2023*, pp. 209–214, 2023, doi: 10.1109/ICITCOM60176.2023.10442970.

[18] Y. Liu and Y. Liu, "LLM-Driven Sentiment Analysis in MD & A : A Multi-Agent Framework for Corporate Misconduct Prediction," *System*, vol. 13, no. 10, pp. 1–26, 2025, doi: https://doi.org/10.3390/systems13100839.

[19] Ö. Aydin, E. Karaarslan, and F. Safa ERENAY, "Generative AI in Academic Writing: A Comparison of DeepSeek, Qwen, ChatGPT, Gemini, Llama, Mistral, and Gemma," 2025.

[20] W. Etaiwi and B. Alhijawi, "Comparative Evaluation of ChatGPT and DeepSeek Across Key NLP Tasks: Strengths, Weaknesses, and Domain-Specific Performance," *Array*, vol. 27, no. August, p. 100478, 2025, doi: 10.1016/j.array.2025.100478.

[21] L. Xiong *et al.*, "DeepSeek: Paradigm Shifts and Technical Evolution in Large AI Models," *IEEE/CAA J. Autom. Sin.*, vol. 12, no. 5, pp. 841–858, 2025, doi: 10.1109/JAS.2025.125495.

[22] D. Wang *et al.*, "Tokenization Matters! Degrading Large Language Models through Challenging Their Tokenization," pp. 1–19, 2024, [Online]. Available: http://arxiv.org/abs/2405.17067

[23] R. Sutoyo, S. Achmad, A. Chowanda, E. W. Andangsari, and S. M. Isa, "PRDECT-ID: Indonesian product reviews dataset for emotions classification tasks," *Data Br.*, vol. 44, p. 108554, Oct. 2022, doi: 10.1016/J.DIB.2022.108554.

[24] H. Akbar, D. Aryani, M. K. Mohammed Al-shammari, and M. B. Ulum, "Sentiment Analysis for E-Commerce Product Reviews Based on Feature Fusion and Bidirectional Long Short-Term Memory," *J. Tek. Inform.*, vol. 5, no. 5, pp. 1385–1391, 2024, doi: 10.52436/1.jutif.2024.5.5.2675.

[25] L. Liu, J. Meng, and Y. Yang, "LLM technologies and information search," *J. Econ. Technol.*, vol. 2, no. November, pp. 269–277, 2024, doi: 10.1016/j.ject.2024.08.007.

[26] H. Tohir, N. Merlina, and M. Haris, "Utilizing Retrieval-Augmented Generation in Large Language Models To Enhance Indonesian Language Nlp," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 10, no. 2, pp. 352–360, 2024, doi: 10.33480/jitk.v10i2.5916.

[27] W. Q. Leong, J. G. Ngui, Y. Susanto, H. Rengarajan, K. Sarveswaran, and W. C. Tjhi, "BHASA: A Holistic Southeast Asian Linguistic and Cultural Evaluation Suite for Large Language Models," 2023, [Online]. Available: http://arxiv.org/abs/2309.06085

[28] Z. Zhou *et al.*, "A Survey on Efficient Inference for Large Language Models," pp. 1–36, 2024, [Online]. Available: http://arxiv.org/abs/2404.14294

[29] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLORA: Efficient Finetuning of Quantized LLMs," *Adv. Neural Inf. Process. Syst.*, vol. 36, 2023.