Sentiment Analysis of Noisy Malay Text using a Large Language Model

Khairul Imran Khalip
Centre for Mathematical Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah

Ku Muhammad Naim Ku Khalif
Centre for Mathematical Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah

Mohd Khairul Bazli Mohd Aziz
Centre for Mathematical Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah

Alexander Gegov
Faculty of Technology, University of Portsmouth

出版情報：Proceedings of International Exchange and Innovation Conference on Engineering & Sciences (IEICES). 11, pp.1904-1909, 2025-10-30. International Exchange and Innovation Conference on Engineering & Sciences
バージョン：

Sentiment Analysis of Noisy Malay Text using a Large Language Model

Khairul Imran Khalip[1], Ku Muhammad Naim Ku Khalif[1, 2], Mohd Khairul Bazli Mohd Aziz[1, 2], Alexander Gegov[3, 4]
[1]Centre for Mathematical Sciences,
Universiti Malaysia Pahang Al-Sultan Abdullah, 26300, Kuantan, Pahang, Malaysia,
[2]Centre for Artificial Intelligence & Data Science,
Universiti Malaysia Pahang Al-Sultan Abdullah, 26300, Kuantan, Pahang, Malaysia,
[4]Faculty of Technology, University of Portsmouth, Portsmouth PO1 3HE, United Kingdom,
[5]English Faculty of Engineering, Technical University of Sofia, 1756 Sofia, Bulgaria
kunaim@umpsa.edu.my

Abstract: *Due to the informality of social media, Malay user-generated content sentiment analysis is difficult. Existing methods struggle to capture cultural and contextual details. This study proposes publishing an open-source annotated dataset, fine-tuning an open-source large language model (LLM), and using an open-source chatbot interface to create a robust sentiment analysis model for noisy Malay text. The research addresses three main issues: lack of labelled Malay social media data, insufficient generic Malay language models, and lack of practical sentiment analysis tools. Its three goals are to create a diverse dataset with accurate sentiment labels, parameter-efficiently fine-tune an LLM, and export the model for an interactive chatbot. The process involves collecting social media data using Contextual Lexical Adaptation, preprocessing and analysing it, fine-tuning the TinyLlama LLM using LoRA, and comparing it to traditional models. Real-world applications, such as sentiment analysis of Malaysian tweets, will be shown using a locally deployed chatbot interface for fine-tuned model inference. This study lays the groundwork for practical sentiment analysis, benefiting businesses, researchers, and politicians seeking data-driven insights. This research aims to revolutionise open-source Malay sentiment analysis by addressing current limitations through an integrated approach.*

Keywords: Sentiment Analysis; Large Language Model;Fine-Tuning;

1. INTRODUCTION

The domain of Artificial Intelligence (AI) has experienced substantial progress with the emergence of robust language models such as OpenAI's Generative Pre-trained Transformer 4 (GPT-4) and Google Gemini. These models have expanded the range of Natural Language Processing (NLP) applications, encompassing text generation and translation. Nevertheless, these models, including the proprietary GPT-4, primarily concentrate on the English language, resulting in a significant deficiency in NLP tasks for non-English languages such as Malay.

This disparity is especially pronounced in sentiment analysis, an essential NLP application that evaluates opinions, sentiments, and emotions from natural language text to autonomously extract the underlying sentiments [1]. User Generated Content (UGC), including blog comments and social media posts in the Malay language, poses distinct challenges for sentiment analysis due to their frequent deviation from grammatically correct syntax and their inherent noise and ambiguity, complicating processing by standard NLP models [2]. In Malaysia, user-generated content frequently amalgamates Malay, English, and Chinese, replete with colloquialisms, slang, and emoticons.

Two prevalent methods for performing sentiment analysis are lexicon-based and machine learning (ML) approaches. This study will concentrate on the latter. Machine learning has gained significant popularity for sentiment analysis owing to its capacity to learn from data and enhance its accuracy progressively. Access to extensive labelled data allows machine learning algorithms to discern patterns and correlations, facilitating the precise classification of sentiment in novel text data. Go et al. (2009) demonstrated that machine learning techniques, including Naïve Bayes (NB) and Support Vector Machines (SVM), effectively classified the sentiment of Twitter messages as positive or negative through distant supervision. Training data comprising tweets with emoticons served as imprecise labels to attain an accuracy exceeding 80% [3].

Large Language Models (LLMs) employing self-attention mechanisms have demonstrated exceptional efficacy in comprehending noisy text, including User-Generated Content (UGC).

The intricate relationships between words and their context facilitate a deeper comprehension of sentiment conveyed through informal language, slang, and emoticons. BERT [4] and RoBERTa [5] are two prominent large language models that have attained superior performance in numerous natural language processing tasks, such as sentiment analysis. Nonetheless, the majority of LLMs are predominantly trained on English text, rendering them less efficient for non-English languages such as Malay. Consequently, there is a necessity for open-source models specifically refined for Malay sentiment analysis. Our research seeks to fill this gap by utilising a variant of Meta's LLaMA2 known as TinyLlama, an open-source and customisable Large Language Model (LLM). We concentrate on optimising TinyLlama for Malay sentiment analysis, specifically addressing the informal and diverse expressions common in user-generated content, such as social media texts.

The proposed methodology entails utilising a labelled dataset of user-generated content in the Malay language to refine TinyLlama's pre-trained weights. Subsequently, we assess the model's efficacy using a distinct test set of Malay UGC texts. Additionally, we employed an open-source chatbot interface to illustrate the practical implementation of the fine-tuned Large Language Model (LLM). In this study, the objective is to illustrate the efficacy of employing an open-source LLM for Malay sentiment analysis and to enhance inclusivity and diversity in NLP research and applications.

This section presents the origins of concept and issues related to sentiment analaysis for noise Malays text. Section 2 reviews the related works of sentiment analysis of malay text using LLM. The proposed large language model framework for noise malay text are shown in Section 3. Section 4 illustrates the results discussion from the findings and making some important points. Section 5 concludes the paper.

2. LITERATURE REVIEW

2.1 Machine Learning Techniques for Malays Sentiment Analysis

Sentiment analysis uses various approaches to interpret emotions in text, mainly categorized into lexicon-based, machine learning, and hybrid methods. Lexicon-based approaches rely on predefined sentiment dictionaries but often fail to capture context and nuanced meanings. They assess sentiment based on keyword presence, which may misrepresent the intended tone.

Machine learning methods offer a more adaptive approach, training models on large datasets to recognize complex text features. Azmi et al. (2022) showed that models like SVM and Naive Bayes effectively classified Malay sentiment using 33,376 MySejahtera feedback entries, with SVM and Vader achieving high accuracy. However, traditional ML still struggles with deeper contextual understanding [6].

Hybrid models combine the strengths of both approaches. Amirah et al. (2022) integrated deep learning with Malay-specific lexicons, outperforming SVM, NB, and RF models. Yet, deep learning requires large labeled datasets and lacks explainability [7].

Sadanandan et al. (2016) introduced a hybrid model merging a structured knowledge base with machine learning to enhance Malay sentiment analysis. This model addressed challenges such as negation, metaphors, and slang, achieving high accuracy by incorporating deeper semantic insights. Figure 1 depicts a negative sentiment concept, “Cost-Rise,” linked to the broader “Cost” [2].

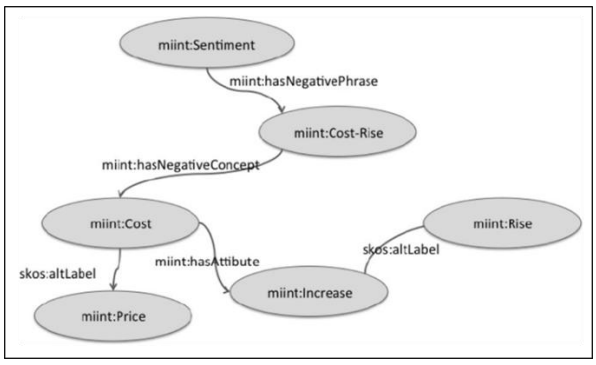

Fig 1. Sentiment knowledge base sample. Adapted from A. A. Sadanandan et al. (2016).

2.2 Large Language Models (LLMs)

The development of large language models (LLMs) like OpenAI’s GPT series marks a major milestone in AI. These models generate human-like text by predicting word sequences based on patterns learned from vast text datasets. Their “knowledge” is stored in parameters—numerical values refined during training to improve accuracy.

Training involves exposing the model to large text corpora and adjusting its parameters as it learns to predict hidden words, much like learning through reading and practice. Fine-tuning may follow, adapting the model to specific tasks or styles. A key aspect of LLMs is their ability to memorise and reproduce parts of their training data, demonstrating both pattern recognition and retention [8].

Since the first GPT in 2018, OpenAI has released several improved versions: GPT-2 (2019), GPT-3 (2020), GPT-3.5 (2022), and GPT-4 (2023), each with more parameters and better performance. These models are trained on diverse datasets like WebText, sourced from 45 million Reddit links, and Common Crawl, a broad snapshot of internet content [9].

2.3 LLMs for Sentiment Analysis in Malays Language

The rise of large language models (LLMs), such as OpenAI’s GPT series, marks a major milestone in AI. These models generate human-like text by learning from vast and diverse datasets. For sentiment analysis in Malay, LLMs like GPT-4 offer a significant leap beyond traditional machine learning, particularly in handling context—a critical aspect of sentiment in Malay.

LLMs excel in capturing long-range dependencies, idiomatic expressions, and varied writing styles, making them well-suited for the linguistic diversity of Malay. Unlike conventional models, which often need extensive fine-tuning, LLMs generalize effectively across different contexts, improving accuracy in sentiment detection.

Their advanced word embeddings enhance semantic understanding, especially in detecting nuanced sentiments across extended texts. However, a notable challenge remains in interpreting sarcasm, a common feature in online Malay content. Expressions like terpaling pakar can be genuine or sarcastic, and distinguishing between the two is difficult even for humans—let alone machines.

While LLMs outperform earlier models in many respects, sarcasm remains a shared limitation. Despite this, their use in Malay sentiment analysis represents significant progress. The next section will explore fine-tuning strategies for LLMs.

LLMs predict the most likely word sequences based on patterns learned during training. Their "parameters" represent internalized linguistic knowledge. Training involves exposing the model to large text corpora and gradually refining predictions, akin to learning through reading and practice. Post-training fine-tuning allows

specialization in specific tasks or domains.
Memorization is another key feature—LLMs can reproduce chunks of their training data, demonstrating both pattern recognition and memory [8].
Since 2018, OpenAI's GPT models have evolved rapidly: GPT-2 (2019), GPT-3 (2020), GPT-3.5 (2022), and GPT-4 (2023), each improving in scale and capability. Their training data includes sources like WebText (from Reddit links) and Common Crawl, ensuring broad coverage and high-quality content for language learning [9].

3. PROPOSED METHODOLOGY

In this study, there are eight main phases in this study. Details of each phase are provided.

Phase 1: Business Understanding

This research aims to develop a robust sentiment analysis model for noisy Malay text from social media, blogs, and forums. It classifies sentiments as positive or negative, handling challenges like slang, multilingualism, and emojis. The model supports businesses in understanding customer feedback, improving marketing, and tracking brand perception. In finance, it aids in gauging market sentiment to guide investment decisions. Politically, it helps understand public opinion, informing strategies and policy communication.

Phase 2: Analytical Method

A binary sentiment classification approach (positive/negative) is used for its clarity and simplicity. The Malaysian-TinyLlama-1.1B-16k-instructions model will be fine-tuned, with performance benchmarked against traditional machine learning models to validate effectiveness.

Phase 3: Data Specifications

Data includes informal Malay text with slang, code-switching, and emojis from social media and blogs. Only positive and negative sentiments will be labelled initially due to resource constraints. Data is stored in JSONL format with six fields: id, text, sentiment, labeller_id, collected_datetime, and labelled_datetime.

Phase 4: Data Acquisition

Data is collected from X (Twitter), Facebook, and blogs, using two lexicons (772 negative and 265 positive words) contributed by Malay social media users. Only content with at least three keywords is included. A method called Contextual Lexical Adaptation (CLA) is introduced to collect sentiment-rich data and reduce mislabeling. CLA also supports contrastive learning, helping the model better understand sentiment expressions in various contexts and slang, improving semantic robustness during fine-tuning.

Table 1. Table captions should appear above the tables. If the heading is no longer than two lines or consists of two sentences, it should not end with a full stop

Negative List	Positive List
butoh	*subhanallah*
mak ko	*alhamdulillah*
hijau	
pondan	*penyayang*
haram jadah	*terima kasih*
sakau	*terbaekkk*
munafik	*rahmat*
Dapig	*berdedikasi*

Phase 5: Data Understanding

Comprehensive analysis will be conducted to understand the dataset. Table 2 summarises data types, examples, and column descriptions, guiding the next phase.

Table 2. Specifications adopted for simulated inverter

Attribute	Data Type	Example	Description
id	Integer	80741	A unique identifier for each text entry.
text	*String*	*Kepala bapak dia lah suka suka nak potong barisan!*	The actual text content in the Malay language.
sentiment	String	Negative	The sentiment or emotional tone conveyed in the text.
labeller_id	Integer	1	The identifier of the labeller who assigned the sentiment label to the text.
collected_datetime	Datetime	2022-02-15T08:08:13.568662Z	The timestamp indicates when the text was originally gathered.
labelled_datetime	Datetime	2022-02-15T08:08:13.568662Z	The timestamp indicates when the sentiment label was assigned to the respective text.

Phase 6: Data Preparation

Minimal pre-processing is applied to preserve linguistic features like emojis, abbreviations, and capitalisation, which may convey sentiment. Tokenisation, duplicate removal, and handling of missing labels will be performed. LLMs like TinyLlama can inherently manage informal text without extensive pre-processing. Python (with Pandas, PyTorch, Transformers, and Unsloth) is used to load and prepare the dataset.

Phase 7: Modelling

TinyLlama will be fine-tuned using LoRA, which enables efficient model adaptation with fewer trainable parameters. Key hyperparameters include:
r (rank size): balances model simplicity and performance,
num_epochs: controls training cycles,
learning_rate: affects convergence speed and accuracy.

Figure 2 illustrates the evolution of our model and its classification within the Llama2 family.

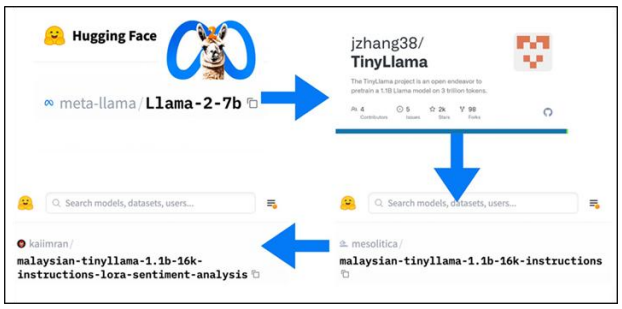

Fig 2. Evolutionary progression of the model.

Phase 8: Model Evaluation

Model performance will be assessed using accuracy, precision, recall, and F1-score. Comparisons with Naive Bayes and SVM will highlight TinyLlama's advantages over traditional methods.

4. RESULT AND DISCUSSION

Exploratory Data Analysis

An exploratory data analysis was initially performed on the labelled dataset gathered for this study. Figure 3 depicts the sentiment class distribution.

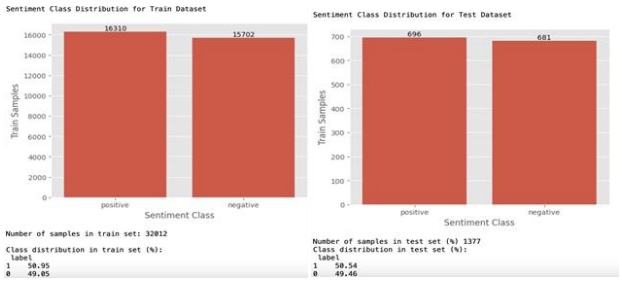

Fig 3. Sentiment class distribution.

Figure 3 shows bar plots of sentiment distribution, highlighting the balance of positive and negative tweets in both training and testing datasets—crucial for building an unbiased model. An imbalanced dataset may cause the model to favor the dominant class. The figure also includes an analysis of tweet lengths based on character count.

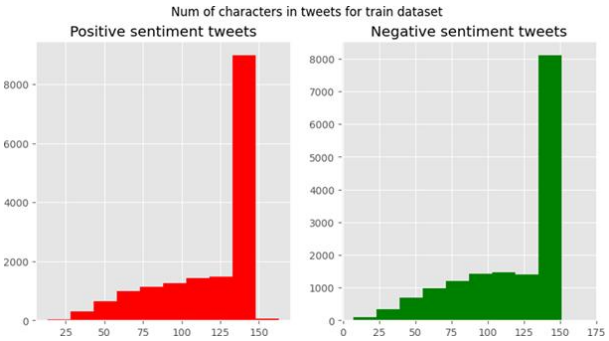

Fig 4. Number of characters of positive and negative sentiment tweets.

Based on Figure 4, positive tweets are generally shorter than negative ones, which are often longer due to complaints or criticisms. Word clouds for the train and test datasets were also generated to highlight the most frequent words.

Fig 5. Word cloud for train and testing datasets.

Figure 5 shows similar word clouds for both train and test sets, dominated by URLs (e.g., https, t.co) and stop words. This is expected, as we skip cleaning steps like URL and stop word removal to align with real-world LLM chatbot usage.

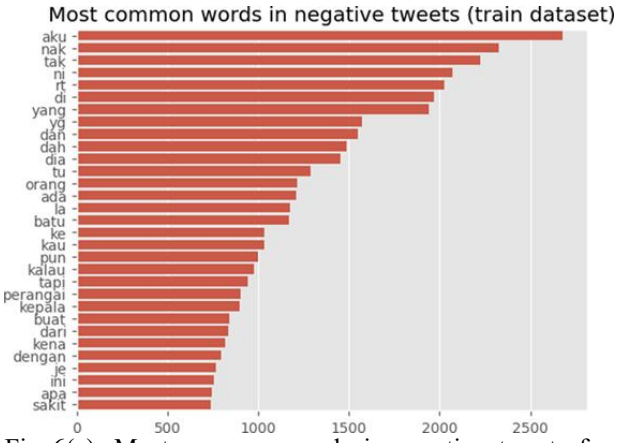

Fig 6(a). Most common words in negative tweets for train dataset.

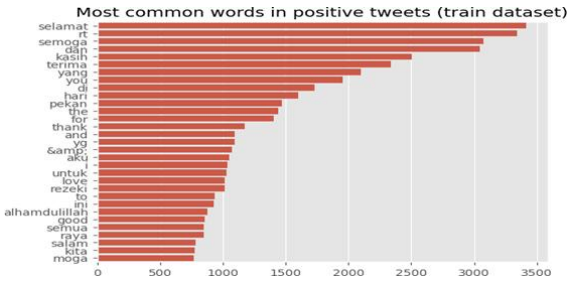

Fig 6(b). Most common words in positive tweets (train dataset).

Based on Figures 6(a) and 6(b), words like *selamat* and *semoga* are common in positive tweets, while negative tweets mostly contain stop words with no dominant swear words, indicating high variation in their usage.

Model Fine-Tuning

Our base model is Malaysian-Tinyllama-1.1B-16k-instructions, a fine-tuned variant of TinyLlama trained on diverse Malaysian texts by Mesolitica. Open-sourced on Hugging Face, it offers strong Malay language understanding for tasks like sentiment analysis on noisy UGC. Choosing this compact model helps reduce project complexity and cost while maintaining good performance. If successful, we may scale to larger models like LLaMA2.

The model was fine-tuned using our labelled dataset (https://huggingface.co/datasets/kaiimran/malaysia-tweets-sentiment) via the Hugging Face Transformers library. We adopted the Low-Rank Adaptation (LoRA) method to update only a small subset (1–10%) of parameters, making training more efficient.

The training and test datasets were loaded and tokenised. Text and sentiment labels were inserted into an instruction template, with an end-of-sentence token appended to prevent infinite text generation.
Figure 7 illustrates the addition of a new attribute, 'input', which combines the values from the 'text' and 'label_text_malay' attributes using the instruction prompt template.

```
from datasets import load_dataset

# Load our dataset from Hugging Face
train_dataset = load_dataset("kaiimran/malaysia-tweets-sentiment", split="train")
test_dataset = load_dataset("kaiimran/malaysia-tweets-sentiment", split="test")

alpaca_prompt = """
### Teks: {}
---
Kenal pasti sama ada teks ini secara keseluruhannya mengandungi sentimen positif atau negatif.
Jawab dengan hanya satu perkataan: "positif" atau "negatif".

Sentimen:
{}"""

EOS_TOKEN = tokenizer.eos_token
def formatting_prompts_func(examples):
    # from pdb import set_trace
    # set_trace()
    inputs       = examples["text"]
    outputs      = examples["label_text_malay"]
    texts = []
    for input, output in zip(inputs, outputs):
        # Must add EOS_TOKEN, otherwise your generation will go on forever!
        text = alpaca_prompt.format(input, output) + EOS_TOKEN
        texts.append(text)
    return { "input" : texts, }

train_dataset = train_dataset.map(formatting_prompts_func, batched=True)
test_dataset = test_dataset.map(formatting_prompts_func, batched=True)

{'text': 'Drug smuggler should die. Sibuk negara org.. padahal negara sendiri pun ada hukuman gantung sampai mati..
gilaa',
 'label': 0,
 'label_text': 'negative',
 'label_text_malay': 'negatif',
 'input': 'Lakukan analisis sentimen bagi teks di dalam tanda sempang berikut.\n——\n### Teks: Drug smuggler should
die. Sibuk negara org.. padahal negara sendiri pun ada hukuman gantung sampai mati..gilaa\n——\nKenal pasti sama ad
a teks ini secara keseluruhannya mengandungi sentimen positif atau negatif.\nJawab dengan hanya satu perkataan: "po
sitif" atau "negatif".\n\nSentimen:\nnegatif</s>'}
```

Fig 7. Tokenzing text in a prompt template.

Evaluation and Comparison of Models
This part presents the evaluation results of the fine-tuned Malaysian-TinyLlama-1.1B model for Malay sentiment analysis. The model was trained and tested using datasets with balanced class distributions. The training set comprises 32,012 samples, while the test set includes 1,377 samples.

MultinomialNB Classification Report:

	precision	recall	f1-score	support
0	0.9122	0.9765	0.9433	681
1	0.9753	0.9080	0.9405	696
accuracy			0.9419	1377
macro avg	0.9438	0.9423	0.9419	1377
weighted avg	0.9441	0.9419	0.9419	1377

SVC Classification Report:

	precision	recall	f1-score	support
0	0.9487	0.9780	0.9631	681
1	0.9778	0.9483	0.9628	696
accuracy			0.9630	1377
macro avg	0.9632	0.9631	0.9630	1377
weighted avg	0.9634	0.9630	0.9630	1377

Fig 8. Word cloud for train and testing datasets.

Figure 9 shows that MultinomialNB achieves 94.19% accuracy, close to the fine-tuned TinyLlama's 94.55%, indicating strong performance. SVC outperforms both with 96.30% accuracy, making it the best on the test set. Overall, the Malaysian-TinyLlama-1.1B model performs competitively with high precision, recall, and F1-score, though SVC holds a slight advantage in accuracy.

5. CONCLUSION

This study successfully fulfilled its three primary objectives, marking a significant advancement in sentiment analysis for noisy Malay text and presenting a viable open-source alternative to proprietary models like GPT-4. By developing a richly annotated and diverse dataset of Malay user-generated content (UGC)—inclusive of emojis, slang, code-switching, and informal expressions—the research addresses a long-standing gap in resources for underrepresented languages. This dataset not only improves accessibility for researchers working in low-resource NLP but also establishes a new benchmark in sentiment analysis complexity, particularly for informal and multilingual digital content. The use of Low-Rank Adaptation (LoRA) to fine-tune an open-source large language model (a variant of TinyLlama) showcased the method's efficiency and effectiveness. The model demonstrated strong accuracy and robustness across sentiment classification tasks, standing on par with or outperforming traditional machine learning baselines. The integration of contrastive learning elements further enhanced the model's ability to interpret contextual sentiment cues in diverse linguistic expressions.
A key innovation of this study is the development of an interactive chatbot interface, which operationalizes the research by enabling non-technical users to access and benefit from advanced sentiment analysis tools. This interface not only demonstrates the model's practical utility but also promotes public engagement with AI-driven language tools in Malay, supporting broader digital inclusion.
Beyond its technical contributions, this research carries meaningful societal and industrial implications. Businesses can leverage the system to monitor consumer sentiment more effectively; policymakers can use it to gauge public response to policy shifts; and social researchers can explore trends in public discourse. The ability to accurately process noisy, informal Malay language at scale empowers decision-makers across sectors with more nuanced, real-time insights.
Despite these contributions, certain limitations remain. The current scope is limited to binary sentiment classification (positive/negative), with neutral and mixed sentiments to be explored in future iterations. Additionally, the dataset, though diverse, may still underrepresent regional dialects or domain-specific terminology. Addressing these areas will further improve the model's generalisability and robustness.
Looking ahead, future research can expand the sentiment classification to multi-class or aspect-based sentiment analysis, incorporate emotion detection, and explore domain-specific fine-tuning for sectors like healthcare, politics, or education. The dataset itself can be scaled and enriched to capture broader linguistic nuances across demographics and regions. Moreover, incorporating multimodal inputs—such as text combined with images or voice—could extend the model's utility in more interactive digital environments.
In conclusion, this study not only establishes a solid framework for Malay sentiment analysis using open-source LLMs but also contributes valuable resources, tools, and methodologies for the NLP community. It exemplifies how responsible AI development can empower linguistic diversity, promote open research, and bridge the digital divide in underrepresented

languages. Through this foundation, future work can build more inclusive, adaptive, and impactful AI solutions for real-world language challenges.

6. REFERENCES

[1] Lombardo, G., Fornacciari, P., Mordonini, M., Sani, L., & Tomaiuolo, M. (2019). A combined approach for the analysis of support groups on Facebook: The case of patients of hidradenitis suppurativa. Multimedia Tools and Applications, 78(3), 3321-3339

[2] Sadanandan, A. A., Osman, N. A., Saifuddin, H., Ahamad, M. K., Pham, D. N., & Hoe, H. (2016). Improving accuracy in sentiment analysis for Malay language. In Proceedings of the 4th International Conference on Artificial Intelligence and Computational Science (pp. 28-29)

[3] Go, A., Bhayani, R., & Huang, L. (2009). Twitter Sentiment Classification Using Distant Supervision. CS224N Project Report, Stanford, 1-12

[4] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171-4186. Minneapolis, MN: Association for Computational Linguistics

[5] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692

[6] Azmi, P. A. R., Abidin, A. W. Z., Mutalib, S., Zawawi, I. S. M., & Halim, S. A. (2022). Sentiment Analysis on MySejahtera Application during COVID-19 Pandemic. In 2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS) (pp. 215-220). IPOH, Malaysia. https://doi.org/10.1109/AiDAS56890.2022.9918748

[7] Amirah, N., Yusoff, M., & Kassim, M. (2022). Hybrid Machine Learning Methods with Malay Lexicon for Public Polarity Opinion on Water Related Issue. In 2022 IEEE International Conference in Power Engineering Application (ICPEA) (pp. 1-5). Shah Alam, Malaysia. https://doi.org/10.1109/ICPEA53519.2022.9744713

[8] Van den Burg, G. J. J., & Williams, C. K. I. (2021). On Memorization in Probabilistic Deep Generative Models. Neural Information Processing Systems.

[9] OpenAI. (2019, November). GPT-2 Model Card. GitHub. https://github.com/openai/gpt-2/blob/master/model_card.md