Multi-Lingual Malaysian Embedding: Leveraging Large Language Models for Semantic Representations

Husein Zolkepli[*] Aisyah Razak[†] Kamarul Adha[‡] Ariff Nazhan[§]

Abstract

In this work, we present a comprehensive exploration of finetuning Malaysian language models, specifically Llama2 and Mistral, on embedding tasks involving negative and positive pairs. We release two distinct models tailored for Semantic Similarity and Retrieval-Augmented Generation (RAG).

For Semantic Similarity, our 600 million parameter Llama2 model outperforms OpenAI text-embedding-ada-002 across all recall@k metrics for b.cari.com.my, c.cari.com.my, Malay news, and Malaysian Twitter test sets.

In the realm of RAG models, our approach proves competitive with OpenAI text-embedding-ada-002 in the Malaysian context. Notably, our 2 billion parameter Llama2 model achieves superior Recall@5, Recall@10 for the "Melayu" keyword research papers dataset and excels in Recall@3, Recall@5, and Recall@10 for the lom.agc.gov.my dataset.

These findings underscore the effectiveness of our finetuning strategy and highlight the performance gains in both Semantic Similarity and RAG tasks.

All models released at HuggingFace Mesolitica Malaysian Embedding Collection.

1 Introduction

In the wake of the release of ChatGPT, the landscape of conversational AI has seen a surge in the development of chatbots armed with proprietary knowledge bases. The success of these chatbots hinges on their ability to perform semantic search effectively—essentially, retrieving the most relevant information from a knowledge base to provide accurate responses to user queries. This functionality relies on a sophisticated embedding model capable of capturing and understanding the nuances of language.

For English-based applications, companies have readily embraced the closed-source OpenAI text-embedding-ada-002 model as a turnkey solution for robust embedding capabilities. However, when it comes to the Malay language, the performance of such out-of-the-box solutions falls short. The intricacies of Malay linguistics, along with its unique semantic structure, pose challenges that generic models struggle to overcome.

In recognition of this disparity, our research endeavors to address the existing gap in the provision of effective embedding models tailored specifically for the Malay language. Rather than relying on closed-source alternatives, our approach seeks to introduce an open-source solution that not only caters to the linguistic intricacies of Malay but also outperforms existing models in the context of

[*]husein@mesolitica.com
[†]aisyahrazak171@gmail.com
[‡]kamarul.adha360@gmail.com
[§]ariffnzhn@gmail.com

Technical Report.

semantic search. By doing so, we aim to propel the field of natural language processing forward, fostering innovation and accessibility for Malay language applications in the realm of conversational AI and knowledge retrieval.

- **Hard mining embedding dataset:** We utilize OpenAI text-embedding-ada-002 and bge-large-en [1] from Beijing Academy of Artificial Intelligence as base models to convert Malaysian texts into embedding representations. Subsequently, we employ hard mining techniques to refine and optimize these embeddings, enhancing their quality and relevance. This approach aims to extract more meaningful semantic information from the original texts, ensuring our embeddings align seamlessly with the nuances of the Malay language for improved performance in various applications.
- **Synthetic RAG dataset:** We created positive and negative pairs using synthetic QA dataset from Malaysian Mistral [2]. These pairs help enhance semantic retrieval, enabling the model to discern intricate connections between contexts and questions. Positive pairs highlight correct retrievals, while negative pairs offer learning opportunities for refining the model's comprehension. This approach enriches the dataset and contributes to developing a more context-aware embedding model for a nuanced understanding of the Malaysian language context.
- **Finetuned Large Language Model:** We fine-tuned Malaysian Mistral [2] models with 191M and 349M parameters, along with Malaysian Llama2 models of 600M, 1B, and 2B parameters using a contrastive loss. To cater to different embedding needs, we extracted the initial N layers of these models and continued pretraining, creating smaller models customized for specific embedding tasks. This approach allows for adaptability and optimal performance in various scenarios.

2 Hard Mining Dataset Procedure

2.1 Converting to Embedding representation

2.1.1 OpenAI text-embedding-ada-002 base model

The utilization of OpenAI's text-embedding-ada-002 plays a pivotal role in transforming samples sourced from diverse platforms, including b.cari.com.my, carigold, Malaysian Facebook posts, Lowyat, Malaysian news, and Malaysian Twitter, into rich and meaningful embedding representations. The text-embedding-ada-002 model serves as the cornerstone for capturing the semantic essence of these Malaysian texts, offering a standardized and condensed representation that encapsulates the contextual information present in the original content.

These initial embedding representations serve as the baseline dataset, laying the foundation for our subsequent hard mining process. The diverse array of sources ensures that our baseline dataset is comprehensive, capturing the linguistic nuances and variations prevalent across different Malaysian platforms. By employing text-embedding-ada-002, we aim to establish a robust starting point for the subsequent stages of our embedding model development.

As we embark on the hard mining process, the quality and richness of the baseline dataset become crucial. The embeddings generated by text-embedding-ada-002 provide a solid foundation for identifying and isolating challenging cases within the dataset, setting the stage for the refinement and enhancement of our models through targeted training iterations

All embedding representation dataset and implementation published at mesolitica/malaysian-dataset/embedding/ada-002.

2.1.2 bge-large-en base model

Our approach involves addressing the language diversity challenge posed by bge-large-en, which is specifically trained on English datasets. To overcome this limitation, we implemented a multi-step process to enrich our embedding representation. We initiated the process by generating noisy translations for content from b.cari.com.my, c.cari.com.my, Malaysian Reddit, and Malaysian Twitter. Leveraging the capabilities of ChatGPT3.5, we translated these Malaysian texts into standard English. Subsequently, a post-filtering mechanism was applied to refine and enhance the accuracy of the noisy

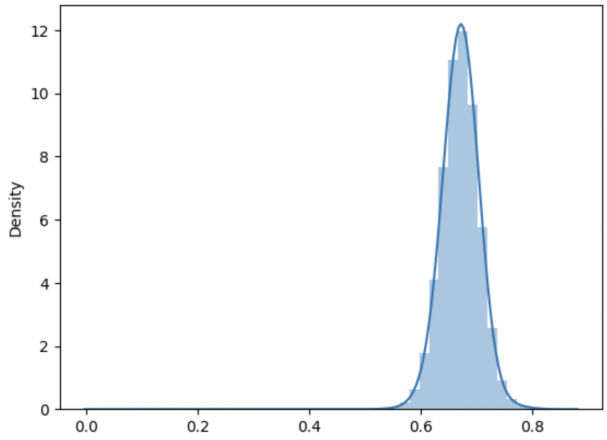
translations, ensuring that the resulting English translations retained the intended semantic meaning of the original Malaysian content.

The translated segments were then used to create an additional set of embedding representations, adding a layer of linguistic diversity to our dataset. This diversification strategy is particularly important to capture the broad spectrum of linguistic nuances present in Malaysian online platforms. By combining the outputs from bge-large-en and OpenAI text-embedding-ada-002, we aimed to create a more comprehensive and representative embedding model. The synergy of these two models contributes to the holistic coverage of the Malaysian language landscape, enhancing the robustness and inclusivity of our embedding representation dataset. This meticulous process ensures that our models can effectively encapsulate the diversity of language expressions found across various Malaysian online sources.

All embedding representation dataset and implementation published at mesolitica/malaysian-dataset/embedding/bge-large-en.

2.2 Hard Mining Procedure

In our analysis, we observed that the distribution of text similarity tends to exhibit a negatively skewed pattern. This indicates that a majority of the generated texts have relatively low similarity scores. Our primary focus lies on the tail-ends of this distribution, particularly on the extreme left-tail, which signifies the most similar texts or what we term as hard positives. On the opposite end, the extreme right-tail represents the most dissimilar texts or hard negatives. By concentrating on these extremes, we aim to capture instances where the texts are either highly similar or distinctly dissimilar, providing a nuanced understanding of the variability within the generated text distribution.



The determination of similarity is guided by the Euclidean distance formula, where lower values indicate greater similarity and higher values signify greater dissimilarity. This rigorous process of hard mining allows us to refine and optimize the embeddings, ensuring that they encapsulate the nuanced semantics of the Malay language. This meticulous approach enhances the quality and relevance of the embeddings, resulting in improved performance across a spectrum of applications.

For a given base text like 'Selamat Tidur mutual ku semua,' the 5% percentile represents the lower end of potential variations. In this instance, the generated results at the 5% percentile include variations such as 'selamat tidur,' 'selamat tidur katanya,' and 'Selamat tidur!!' These variations capture the diversity in possible renditions, showcasing different expressions and styles within the given context. The percentile-based approach allows for a range of outputs, offering insights into the variability of language generation around the base text.

In the pursuit of efficient hard mining for our embedding dataset, we strategically employ the scipy.spatial.KDTree approach. While the prevailing trend involves utilizing Faiss multi-GPUs [3] for mining tasks, our deliberate choice of scipy.spatial.KDTree stems from a consideration of both effectiveness and cost-effectiveness. Leveraging a virtual machine with ample CPU and memory resources proves to be a more economical alternative compared to configurations with multi-GPUs.

The decision to opt for scipy.spatial.KDTree is rooted in our commitment to resource optimization within the confines of budget constraints. By harnessing the power of this spatial data structure, we strike a balance between achieving effective hard mining and ensuring the judicious utilization of available computing resources. This strategic choice aligns with our overarching goal of delivering optimal performance without compromising on cost efficiency.

Below is pseudo Python code to do hard mining,

```
def mining(kd_tree, i, vectors, lower_bound, upper_bound, max_size = 5):
    dist, ind = kd_tree.query(vectors[x], k=len(vectors), workers = 1)

    pos_indices = [k for k in ind[dist <= lower_bound]]
    neg_indices = [k for k in ind[dist > upper_bound]]

    if len(neg_indices) > max_size:
        neg_indices = random.sample(neg_indices,max_size)
    if len(pos_indices) > max_size:
        pos_indices = random.sample(pos_indices,max_size)
```

All hard-mining implementation published at mesolitica/llm-embedding/mining-openai and mesolitica/llm-embedding/mining-bge

3 Retrieval-Augmented Generation Dataset Procedure

Addressing the inherent challenges associated with traditional RAG datasets, where the length of context far surpasses that of user questions, necessitates a nuanced strategy to optimize the performance of our language models. In the typical RAG scenario, contexts may span entire documents or articles, providing a rich source of information, while user queries tend to be concise and specific. This incongruence in length poses a considerable hurdle for many existing embedding models, which may struggle to effectively comprehend and respond to extensive contextual information.

In response to this challenge, we have devised a tailored approach leveraging a synthetic Malaysian Open QA dataset from Malaysian Mistral [2]. This synthetic dataset is curated by supplying contextual information to ChatGPT3.5 and prompting it to generate a comprehensive list of QA pairs based on that context. By employing this method, we ensure that the generated QA pairs align seamlessly with the intricate linguistic nuances of the Malaysian language context, effectively addressing the disparity between the lengths of context and user questions.

Furthermore, we incorporate a robust post-validation mechanism to enhance the overall quality of our dataset. Utilizing a keyword-based validation approach, we meticulously evaluate and accept QA pairs only when the generated answer exhibits a substantial overlap of at least 60% with the original context. This stringent validation criterion serves as a powerful filter, guaranteeing that the generated QA pairs maintain a high degree of relevance and fidelity to the provided context.

This comprehensive approach, combining synthetic dataset generation and meticulous post-validation measures, empowers our models to effectively navigate and understand extended contextual information. The resulting dataset is finely tuned to the specific requirements of RAG tasks on longer texts, contributing significantly to the enhancement of contextual understanding within our language models.

Example of the dataset,

```
 [{'paragraph': 'The Legend of Korra ialah ...',
 'url': 'https://ms.wikipedia.org/wiki?curid=823980',
 'qa': {'qa': [{'question': 'Apakah siri animasi yang ditayangkan di Nickelodeon
     sejak 2012?',
   'answer': 'The Legend of Korra'},
```

```
   {'question': 'Siapakah pencipta siri animasi The Legend of Korra?',
    'answer': 'Bryan Konietzko dan Michael Dante DiMartino'},
   {'question': "Apakah yang dimaksudkan dengan 'bending' dalam siri animasi The
       Legend of Korra?",
    'answer': 'Kekuatan untuk memanipulasi elemen seperti air, bumi, api, atau
        udara'},
   {'question': 'Siapakah Avatar Korra?',
    'answer': 'Pengganti Aang dalam siri sebelumnya yang menghadapi pergolakan
        politik dan roh semangat dalam dunia pemodenan'},
   {'question': 'Apakah kejayaan siri The Legend of Korra?',
    'answer': 'Kejayaan yang kritikal dan komersial dengan jumlah penonton
        tertinggi bagi siri animasi di Amerika Syarikat pada tahun 2012'}]}},
{'paragraph': 'adalah sebuah siri televisyen penstriman ...',
 'url': 'https://ms.wikipedia.org/wiki?curid=1070143',
 'qa': {'qa': [{'question': 'Apakah nama siri televisyen tersebut?',
    'answer': 'Alice in Borderland'},
   {'question': 'Siapakah pengarah siri televisyen tersebut?',
    'answer': 'Shinsuke Sato'},
   {'question': 'Apakah tarikh penayangan perdana siri televisyen tersebut di
       Netflix?',
    'answer': '10 Disember 2020'},
   {'question': 'Apakah ulasan positif yang diterima siri televisyen tersebut?',
    'answer': 'visual, sinematografi, penyuntingan, dan penggunaan grafik
        kekerasan'},
   {'question': 'Berapa musim siri televisyen tersebut?',
    'answer': 'Dua musim'}]}}]
```

The formulation of positive and negative pairs within our dataset intricately contributes to the robustness of our language models. For positive pairs, we draw directly from the actual QA pairs generated based on the provided context. This approach ensures that the positive pairs inherently encapsulate the contextual understanding and thematic relevance of the given information. By anchoring the positive pairs in the real QA generation process, we fortify our dataset with instances that authentically represent the nuanced interplay between context and questions.

The negative pairs play a pivotal role in diversifying the dataset and enhancing the model's ability to discern between relevant and irrelevant information. These negative pairs are sourced from different data points, involving QA pairs generated from alternative contexts. This deliberate introduction of contrasting information encourages the model to refine its discriminatory abilities, discerning not only what constitutes a suitable answer within a given context but also differentiating it from responses generated in dissimilar thematic settings.

By methodically incorporating both positive and negative pairs, our dataset encapsulates a rich spectrum of linguistic intricacies and context-dependent reasoning. This meticulous curation of pairs not only strengthens the contextual understanding of our language models but also fosters a more discerning and adaptive response mechanism, poised to handle a diverse array of user queries within a wide spectrum of contextual scenarios.

Example of generated positive and negative pairs,

```
 {'query': 'The Legend of Korra ...',
'positive_pairs': ['Apakah siri animasi yang ditayangkan di Nickelodeon sejak
    2012?',
 'Siapakah pencipta siri animasi The Legend of Korra?',
 "Apakah yang dimaksudkan dengan 'bending' dalam siri animasi The Legend of
     Korra?"],
'negative_pairs': ['Apakah tarikh penayangan perdana siri televisyen tersebut di
    Netflix?',
 'Bilakah pencabaran terhadap Bill Clinton dimulakan?',
 'Pencabaran terhadap Bill Clinton telah dimulakan pada 8 Oktober 1998.']
 }
```

All synthetic dataset and implementation published at mesolitica/malaysian-dataset/embedding/instructions-pair.

4 Fine-tuning Procedure

4.1 First N hidden layers Continue Pre-training

We are extending the pretraining of our language models by focusing on the initial N layers from the base models. This helps us create a more compact yet insightful representation. Using the same dataset as Malaysian Mistral [2], our approach ensures efficiency. We employ a packing technique with a context length of 32768 to enhance the model's understanding of diverse linguistic contexts.

Our continued pretraining efforts encompass three different parameter sizes: 600 million, 1 billion, and 2 billion. For the 600 million parameter model, we extract information from the first two hidden layers of Malaysian Llama2. Likewise, the 1 billion parameter model draws insights from the first four hidden layers of Malaysian Llama2. Lastly, the 2 billion parameter model leverages knowledge from the first six hidden layers of Malaysian Llama2. This tiered approach allows us to tailor our models to different scales, capturing varying levels of intricacies within the Malaysian language.

During this phase is to maximize the likelihood of the entire sequence,

$$P(x_1, x_2, \ldots, x_T) = \prod_{t=1}^{T} P(x_t | x_1, x_2, \ldots, x_{t-1})$$

We utilized Standard_NC96ads_A100_v4 Azure instance which contains 4x A100 80 GB GPUs. The continue pre-training hyperparameters are detailed below:

Hyperparameter	Value
DeepSpeed	ZeRO-3 Offload
Batch Size	4
Learning Rate	constant 2e-5
Precision	bfloat16

Complete continue pre-training 32768 context length implementation at here.

4.2 Contrastive Fine-tuning

$$\mathcal{L}_{\text{contrastive}}(y, d) = \begin{cases} (1-d)^2 & \text{if } y = 1 \\ \max(d - \alpha, 0)^2 & \text{if } y = 0 \end{cases} \tag{1}$$

where y is the binary label (1 for positive pairs, 0 for negative pairs), d is the dissimilarity score, and α is a margin parameter.

5 Evaluation

We also compared with other models and published the benchmark at mesolitica/malaysian-embedding-leaderboard.

6 Acknowledgement

Special thanks to Malaysia-AI volunteers especially Wan Adzhar Faiq Adzlan, Ammar Azman, M. Amzar, Muhammad Farhan and Syafie Nizam for contributing dataset to train Malaysian Embedding models.

We would like to express our gratitude to NVIDIA Inception for generously providing us with the opportunity to train our model on the Azure cloud. Their support has played a crucial role in the success of our research, enabling us to leverage advanced technologies and computational resources.

We extend our thanks to the wider research community for their valuable insights and collaborative discussions, which have greatly influenced our work. This paper reflects the collective efforts and contributions from both NVIDIA Inception and the broader research community.

7 Conclusion

In conclusion, we have introduced an open-source Malaysian embedding model that exhibits competitive performance, particularly excelling in tasks such as Retrieval-Augmented Generation (RAG) and semantic similarity. This model stands as a viable alternative, eliminating the reliance on closed-source solutions. By offering a publicly accessible and proficient Malaysian embedding model, our contribution aims to foster transparency, accessibility, and innovation within the field, paving the way for diverse applications and further advancements in natural language processing for the Multi-language in Malaysia.

References

[1] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

[2] Husein Zolkepli, Aisyah Razak, Kamarul Adha, and Ariff Nazhan. Large malaysian language model based on mistral for enhanced local language understanding, 2024.

[3] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus, 2017.