Adapting Safe-for-Work Classifier for Malaysian Language Text: Enhancing Alignment in LLM-Ops Framework

Aisyah Razak[*] Ariff Nazhan[†] Kamarul Adha[‡] Wan Adzhar Faiq Adzlan[§]
Mas Aisyah Ahmad[¶] Ammar Azman[‖]

Abstract

As large language models (LLMs) become increasingly integrated into operational workflows (LLM-Ops), there is a pressing need for effective guardrails to ensure safe and aligned interactions, including the ability to detect potentially unsafe or inappropriate content across languages. However, existing safe-for-work classifiers are primarily focused on English text. To address this gap for the Malaysian language, we present a novel safe-for-work text classifier tailored specifically for Malaysian language content. By curating and annotating a first-of-its-kind dataset of Malaysian text spanning multiple content categories, we trained a classification model capable of identifying potentially unsafe material using state-of-the-art natural language processing techniques. This work represents an important step in enabling safer interactions and content filtering to mitigate potential risks and ensure responsible deployment of LLMs. To maximize accessibility and promote further research towards enhancing alignment in LLM-Ops for the Malaysian context, the model is publicly released at malaysia-ai/malaysian-sfw-classifier.

1 Introduction

The AI field, especially natural language processing, [1] has seen remarkable progress with significant breakthroughs like transformer-based architectures [2], multimodality integration to chatbot applications, and reinforcement learning from human feedback. This has led to the rise of open-domain dialogue systems, known as chatbots or conversational agents, which are now increasingly integrated into our daily lives.

Due to the nature of how large language models are trained, using internet data, it is prevalent that there may be harmful contents included. However, as users continues engaging with these chatbots, exposure to harmful and provocative text can have significant adverse effects, impacting individuals' mental well-being, relationships, and emotional state. Therefore, ensuring safe and beneficial interactions has become critically important.

The scarcity of data for identifying not safe for work content, particularly in the Malay language, hinders the advancement of undesired content filtration. Past work such as [1, 3] have laid foundation in terms of AI moderation using large language models, but there is still no not safe for work task in malay language. In this paper, we address this challenge by initiating the data gathering process to

[*] aisyahrazak171@gmail.com
[†] ariffnzhn@gmail.com
[‡] kamarul.adha360@gmail.com
[§] adzhar.faiq@gmail.com
[¶] masaisyahahmad@gmail.com
[‖] ammarakef98@gmail.com

Technical Report.

create a comprehensive dataset of harmful texts. Our methodology involves mining data representative of harmful text categories. Our categorization includes the following labels: pornography, harassment, sexist, racist, religious insult, self-harm, psychiatric or mental illness, and safe for work.

We aim to create a robust classifier tailored for Malaysian language text, enhancing the alignment of our large language model operations framework with safety and ethical standards. This classifier serves as a necessary guardrail within the LLM-Ops framework, providing a cost-effective solution for ensuring safe AI. By systematically identifying and filtering out inappropriate content, this classifier will help create a safe and respectful interaction environment for users.

Furthermore, to the best of our knowledge, there is currently no existing local dataset for the Malaysian language that addresses these specific categories of harmful content. Our work thus represents a pioneering effort in developing and applying this crucial safety measure.

2 Taxonomy

Designing a universal taxonomy for safe for work guardrails is challenging due to the context-dependent nature of language. Below, we outline our taxonomy for safe-for-work categorization, which will guide the application of guardrails to our large language model or chatbot system. Each category is described to clarify the scope and specifics of what constitutes undesired content:

- **Pornography:** Content that includes explicit sexual descriptions, depictions of sexual acts, or nudity intended to arouse sexual interest. This category covers sexually explicit text, adult content descriptions, and any language or media that depicts sexual activity.
- **Harassment:** Content that targets individuals or groups with the intent to demean, intimidate, or threaten. This includes abusive language, threats, stalking, or any form of verbal harassment aimed at causing emotional or psychological distress.
- **Sexist:** Content that promotes discrimination or prejudice based on gender. This includes sexist remarks, derogatory comments about any gender, and language that reinforces harmful gender stereotypes or inequality.
- **Racist:** Content that discriminates or promotes hatred based on race, ethnicity, or nationality. This includes racial slurs, derogatory remarks about ethnic groups, and any language that supports racial superiority or inferiority.
- **Religious Insult:** Content that disrespects or mocks religious beliefs, practices, or figures. This includes blasphemy, offensive jokes about religions, and language intended to insult or offend individuals based on their religious affiliations.
- **Self-Harm:** Content that depicts or encourages self-injurious behavior or suicide. This includes descriptions of self-harm methods, discussions promoting suicide, and any language that glorifies or encourages self-destructive actions.
- **Psychiatric or Mental Illness:** Content that stigmatizes or discriminates against individuals with mental health conditions. This includes derogatory terms for mental health issues, insensitive jokes, signs of mental distress or illness, and any language that trivializes or mocks mental illness.
- **Violence:** Content that promotes or glorifies violence or celebrates the suffering or humiliation of others.
- **Safe for Work:** Content that is appropriate for a professional or public environment, free from explicit, offensive, or discriminatory material. This includes clean language, respectful discourse, and content that does not contain any of the above-mentioned undesired elements.

This taxonomy will help us systematically identify and filter out inappropriate content, where additional redirection can be made to ensure a safe and respectful interaction environment for users.

3 Data Source

The data for this study was collected from various platforms, including social media, public forums, and publicly available datasets. The majority of the data is in the malay language and relevant

to the malay context. By utilizing these comprehensive datasets from multiple sources, we have strengthened the robustness and accuracy of our classification model, enabling it to effectively tackle the challenges of identifying self-harm and sexism in online content.

3.1 Social Media

Data was collected from popular social media platforms such as Twitter and Facebook. Two main approaches were employed to gather relevant data:

1. Keyword-based scraping: a list of keywords associated with explicit content was compiled. These keywords were used to extract tweets from the platform.

2. Profile-based scraping: a list of profiles known for regularly posting NSFW content was curated. Posts from these profiles were then scraped to obtain a more targeted dataset.

The combination of these two scraping methods resulted in a comprehensive and diverse dataset from social media, capturing both keyword-specific content and data from profiles that frequently share explicit material.

3.2 Public Articles

For public articles, we have collected data from various articles and blogs which are b.cari, which hosts a wide range of user-generated content in Malay.

The collected dataset consists of human dialogues extracted from these articles and blog posts. It notably includes some dialogues that contain nsfw content. The inclusion of such dialogues, while potentially controversial, is important to allow the trained classifier to effectively detect explicit content that may realistically occur in open-ended dialogue systems.

3.3 Public Datasets

We have also collected data from publicly available dataset on kaggle such as Kaggle: Suicide and Depression Detection datasets. The dataset is a collection of posts from the "SuicideWatch" and "depression" subreddits of the Reddit platform. This dataset contains a wide range of suicide ideation contexts, providing valuable insights for our research.

We also leverage Explainable Detection of Online Sexism (EDOS) from github and EXIST: sEXism Identification in Social neTwork (EXIST) from web that contain a diverse collection of sexism statements, which have significantly contributed to the success of our classifier in identifying and categorizing such content.

4 Methodology

Supervised text classification requires reliable class labels for training data. However, obtaining these labels can be complex and expensive. Typically, labels are added sequentially by querying an annotator until satisfactory performance is achieved. We introduced an approach that leverages active learning, knowledge distillation of large language models, and text clustering to reduce annotation effort and construct a collection of labeled not safe for work (NSFW) data from our gathered data.

Following figure illustrates the overall flow employed in our methodology to collect NSFW dataset aims to label malaysian dataset for alignment in LLMOps framework.

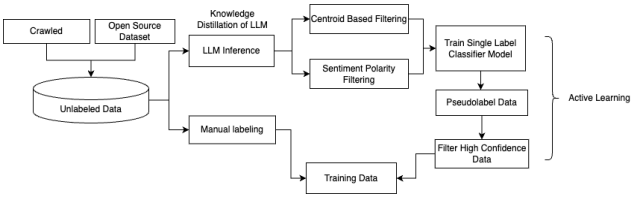


Figure 1: Overall Model Training Framework

4.1 Manual Labeling

Relying on large language models (LLMs) for data labeling can present challenges, such as inaccuracies and biases introduced by the model. To address these issues and enhance the quality of our dataset, we employed manual labeling using Label Studio to complement dataset produced from the label produced by large language models.

Label Studio is an open-source data labeling tool that allows for efficient and precise annotation of data. We used Label Studio to manually label approximately 200 data points, which were then added to our dataset as a baseline.

- **Annotation Process**: Our team meticulously reviewed and annotated the data points to ensure high-quality labels.
- **Baseline Dataset**: The manually labeled data serves as a baseline for our model. By starting with a small, accurately labeled dataset, we provide the model with a strong foundation for learning.

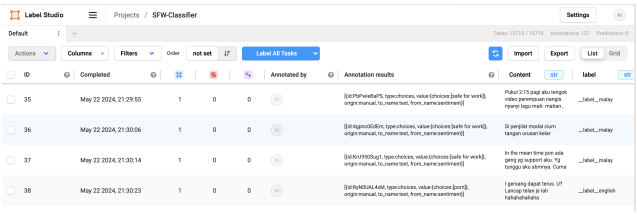


Figure 2: Manual labeling process using Label Studio.

4.2 Knowledge Distillation of Large Language Models

Data annotation using knowledge distillation from a large language model (LLM) involves prompting a teacher LLM to generate a label y for a given input x as the foundational knowledge. This paper primarily explores the zero-shot paradigm, where the LLM receives only the input x and an instruction I without any labeled examples or demonstrations. This straightforward yet effective method of extracting knowledge from teacher LLMs has been widely applied to various tasks and applications. It only requires an input dataset, which is then processed by the LLM to produce the desired label y.

In this annotation stage, we utilized two different large language models: mistral-7b and MaLLaM-small. By leveraging mistral-7b, which excels with English-centric prompts, we translated our scraped data into English to provide better context and achieve reliable outputs from the model. While for MaLLaM [4], we provided inputs from the original malay text as it has better contextual meaning compared to the other models. Below is the prompt we used to generate the labels from the models:

```
text: <text content>

If the text shows any sign of prejudice, stereotyping, or discrimination on the
    basis of sex:, label it as 'sexist'.
If the text shows any sign of content that threatens, incites, glorifies, or
    expresses desire for violence or harm., label it as 'violence'.
If the text shows any sign of prejudice, discrimination, or antagonism by an
    individual, community, or institution against a person or people on the
    basis of their membership of a particular racial or ethnic group, label it
    as 'racist'.
If the text shows any sign of content depicting adult nudity or sexual behavior
    that is pornographic or intended to cause sexual arousal, label it as
    'porn'.
If the text shows any sign of psychiatric or mental illness, label it as
    'psychiatric or mental illness'.
If the text shows any sign of promotion, or otherwise encourage, suicide or
    self-harm, label it as 'self-harm'.
If the text shows any sign of harassment, label it as 'harassment'.
```

```
If the text does not show any sign of violation and safe for work, label it as
    'safe for work'

Only use the label from above choice.

return the result in JSON format {'label', 'explain'}
```

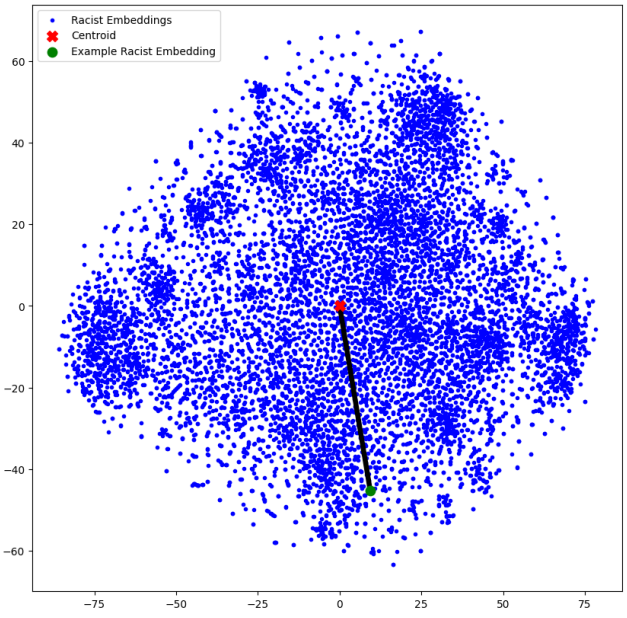
4.3 Centroid Based Filtering

We utilize labeled data obtained from a large language model (LLM) to enhance the quality and consistency of our dataset. First, we compute the centroid of the feature vectors for the labeled data, representing the central point in the feature space. The determination of similarity is guided by the Euclidean distance formula, where lower values indicate greater similarity and higher values signify greater dissimilarity.

Algorithm 1 Centroid-Based Filtering

Input: Labeled text data, LLM for embeddings
Output: Filtered dataset
Generate embeddings for the specific topic labeled by LLM
Compute the centroid of the embeddings
for each embedding **do**
    Calculate the Euclidean distance to the centroid
end for
Observe the distribution of distances
Set a threshold based on the distance distribution
Filter out texts with distances more than the threshold
return Filtered dataset

By measuring the distance of each data point from this centroid, we can identify and filter out data points that are far from the centroid. These distant points are likely to be outliers or less representative of the core topic data distribution. This ensures that the labels encapsulate the nuanced semantics of the specific topic. This filtering process enhances the dataset by retaining data that is more coherent and relevant, thereby improving the performance of the classifier.



4.4 Sentiment Polarity Filtering

Ensuring data label accuracy is crucial for optimal model performance, particularly when dealing with the inherently subjective nature of textual data. To improve the quality of our labeled data, we employed sentiment polarity filtering. This approach involves filtering out data labeled with positive sentiment from our dataset, which was initially produced through knowledge distillation of large language models.

For instance, negative sentences like `Padan muka hang. Tau takut`, categorized under harassment, clearly exhibit negative polarity. In contrast, sentences deemed safe for work typically exhibit positive or neutral polarity. By filtering the dataset based on sentiment polarity, we aim to enhance the dataset's quality, ensuring it includes only the most relevant data.

4.5 Active Learning

We use relevant data from a specific category to train a classifier model. The trained model is then used to classify unlabeled data, a process known as pseudolabeling. The outputs given by the classifier are reviewed before feeding all the unlabeled data into a larger classifier model. This iterative process continues until the desired accuracy is achieved.

The following algorithm outlines the active learning process:

Algorithm 2 Active Learning for Single Label Classifier

Input: Labeled data D_L, Unlabeled data D_U, Classifier model M, Desired accuracy $A_{desired}$
Output: Trained classifier model M
while Accuracy $A <$ $A_{desired}$ **do**
    Train classifier M on D_L
    Predict labels for D_U using M (pseudolabeling)
    Filter high-confidence predictions from D_U to create a new labeled dataset $D_{L_{new}}$
    Update labeled dataset: $D_L \leftarrow D_L \cup D_{L_{new}}$
    Retrain classifier M on the updated D_L
    Manually evaluate the accuracy A of M
end while
return Trained classifier model M

5 Result

In this section, we evaluate the performance of our finetuned model using the training set. I ran evaluation of all the models on the subset of the NSFW dataset. The evaluation metrics used are accuracy, precision, recall, and F1 score. These metrics provide a comprehensive view of the model's performance across different aspects.

5.1 Model Comparison

Table 1: Evaluation on Malaysian NSFW Dataset

Model	Accuracy	Precision	Recall	F1 Score
mesolitica/malaysian-mistral-191M-MLM	0.8768	0.8601	0.8854	0.8714
mesolitica/malaysian-mistral-191M-4096	0.82583	0.81867	0.81657	0.81556
microsoft/debertav3-base	0.26646	0.02961	0.11111	0.04676

Table 1 summarizes the performance metrics for different models. Each model was trained and evaluated with the same test set under the same conditions to ensure a fair comparison.

The results indicate that the `mesolitica/malaysian-mistral-191M-MLM` model outperforms the other models across all metrics. It achieves the highest accuracy of 0.8768, as well as strong precision, recall, and F1 score values of 0.8601, 0.8854, and 0.8714, respectively. This suggests that `mesolitica/malaysian-mistral-191M-MLM` is the most effective model for identifying NSFW content in the Malaysian dataset.

The success of the `mesolitica/malaysian-mistral-191M-MLM` model can be attributed to the implementation of the LLM2Vec [5] approach. LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders, is a simple and efficient solution to transform any decoder-only LLM into a powerful text encoder in an unsupervised fashion using adapters (LoRA), without the need to modify the base models.

The `microsoft/debertav3-base` model has the lowest performance among the compared models, with an accuracy of 0.26646 and significantly lower precision, recall, and F1 score values of 0.02961, 0.11111, and 0.04676, respectively. This model is not suitable for NSFW content detection in the Malaysian dataset.

In summary, the `mesolitica/malaysian-mistral-191M-MLM` model is the most suitable choice for NSFW content detection in the Malaysian context, providing the highest accuracy and consistency across various performance metrics. The `mesolitica/malaysian-mistral-191M-4096` model, while less effective, still maintains a respectable level of performance. The `microsoft/debertav3-base` model, however, does not perform adequately for this task.

5.2 Analysis

To gain more insight into the results of our pseudolabeling process, we performed topic modeling for each respective label. This allowed us to explore the topics present within each label in greater detail.

We employed Latent Dirichlet Allocation (LDA) as our topic modeling technique. LDA is a popular method for discovering hidden topics within a collection of documents. By treating each label as a separate document corpus, we were able to identify the dominant topics within each label category.

Figure 3: Harassment Wordcloud

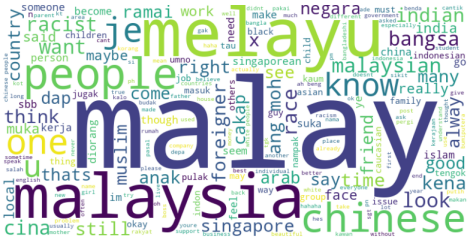


Figure 4: Racist Wordcloud

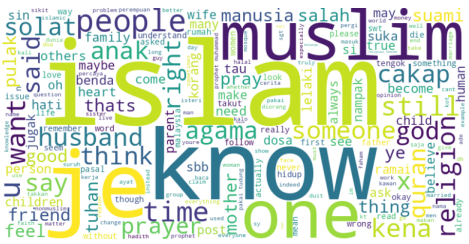


Figure 5: Religion

Figure 6: Self-harm Wordcloud

Figure 7: Porn Wordcloud

Figure 8: Violence Wordcloud

Figure 9: Sexist Wordcloud

Figure 10: Psychiatric or Mental Illness Wordcloud

6 Acknowledgement

We would like to express our gratitude to Mesolitica for providing us with the resources to train our model. 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, our work has presented a novel approach to moderating harmful content in AI safety through the introduction of a Safe For Work (SFW) classifier. By harnessing the power of large language models and data mining techniques, we have constructed a valuable dataset annotated with harmful topics and the first SFW classifier for Malaysian texts. This pioneering research fills a significant gap in the field, as no prior studies have explored this specific approach. Encouraged by our promising results, we leave for future work the refinement of the classifier to better distinguish among varying levels of harmful content and different types of harmful topics. Given the vast amount of user-generated content online, we believe this work represents a significant step forward in AI safety and content moderation.

References

[1] Todor Markov, Chong Zhang, Sandhini Agarwal, Tyna Eloundou, Teddy Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world, 2023.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[3] Huachuan Qiu, Shuai Zhang, Hongliang He, Anqi Li, and Zhenzhong Lan. Facilitating pornographic text detection for open-domain dialogue systems via knowledge distillation of large language models, 2024.

[4] Husein Zolkepli, Aisyah Razak, Kamarul Adha, and Ariff Nazhan. Mallam – malaysia large language model, 2024.

[5] Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. Llm2vec: Large language models are secretly powerful text encoders, 2024.