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

1 Introduction

2 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.

2.1 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.

2.2 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

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

potentially controversial, is important to allow the trained classifier to effectively detect explicit content that may realistically occur in open-ended dialogue systems.

2.3 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.

3 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.

3.1 Knowledge Distillation

```
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’}
```

3.2 Centroid 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. 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 data distribution. This process helps in refining the dataset by retaining data that is more coherent and relevant, thereby improving the classifier's performance.

3.3 Pseudolabeling

We use relevant data of specific category to train a classifier model. The trained model is then will be used to classify unlabeled data, this process is called pseudolabeling. We provisioned the output given by the classifier before feeding all the unlabeled data into bigger classifier model. This process is repeated until we get the desired accuracy.

Below are the example of pseudolabeling process:

4 Result

4.1 Evaluation

5 Acknowledgement

Special thanks to Malaysia-AI volunteers especially Ammar Azman, M. Amzar, Muhammad Farhan, Syafie Nizam, Alif Aiman, Azwan Zuharimi and Haziq Zikry for contributing dataset to train Malaysian Reranker 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.

6 Conclusion

References