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

Aisyah Razak[*] Ariff Nazhan[†]

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.

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 Dataset

3 Finetuning Procedure

4 Evaluate

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.

[*]aisyahrazak171@gmail.com
[†]ariffnzhn@gmail.com

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