LLM-Based Sentiment Analysis, Toxicity Detection, and Detoxification

Tim Christilaw Jacob Nadal COLX 565 UBC MDS-CL

Abstract

In this report, we describe our enhanced LLM-based NLP framework designed for sentiment analysis, toxicity detection, and toxic-to-non-toxic style transfer. Building upon our previous milestone, we replaced OpenAI's GPT-3.5 API with an open-source workflow using UBC-NLP's Toucan model for multilingual translation and IBM Granite for classification and detoxification tasks. We outline our implementation, evaluation metrics, and challenges encountered in this milestone. Our results show competitive performance in sentiment and toxicity detection while introducing a novel detoxification pipeline.

1 Introduction

Natural language processing (NLP) has seen significant advancements with the introduction of large language models (LLMs), enabling tasks such as sentiment analysis, toxicity detection, and text detoxification. In this milestone, we extend our original framework to incorporate multilingual support and a detoxification step, allowing us to rewrite harmful content into a more neutral or constructive form. The main modifications include:

- Using **LangDetect** for fast language detection.
- Incorporating **UBC-NLP/Toucan-base** for multilingual translation.
- Replacing OpenAI's GPT-3.5 API with **IBM Granite** for sentiment classification, toxicity detection, and detoxification.
- Evaluating performance on new test datasets provided in **Milestone-2-toxic-test-solutions.csv** and **Milestone-2-multilingual-sentiment-test-solutions.csv**.

Our work ensures explainability by providing rationales for each classification while incorporating structured error handling. The complete project repository, including the code and additional documentation, is available at: `https://github.ubc.ca/MDS-CL-2024-25/COLX_565_Project_Jacob-Tim`.

2 Framework Setup

Our framework follows a structured pipeline consisting of three primary tasks:

1. **Language Detection and Translation**: We use LangDetect to determine the language of the input text. If the text is not in English, we use the UBC-NLP/Toucan-base model to translate it into English.
2. **Sentiment and Toxicity Classification**: Once in English, the text is analyzed using IBM Granite to determine sentiment (*positive, negative, mixed*) and toxicity (*toxic, non-toxic*).
3. **Detoxification**: If the text is classified as toxic, IBM Granite rewrites the text into a non-toxic alternative while preserving meaning.

The system is designed to be modular, allowing each step to function independently while maintaining structured output integrity.

3 Approach and Implementation

The NLP pipeline is implemented as follows:

- **Language Detection**: We use the LangDetect library to quickly identify the language of input text. If the detected language is not English, we proceed to translation.

- **Translation**: UBC-NLP/Toucan-base is used for many-to-many translation, ensuring accurate conversions into English before further processing.
- **Sentiment and Toxicity Analysis**: IBM Granite is used to classify text sentiment (positive, negative, mixed) and toxicity (toxic, non-toxic). The model provides justifications for its classifications to enhance interpretability.
- **Detoxification**: If a text is flagged as toxic, IBM Granite rewrites the text into a polite version while maintaining its original intent.

The entire pipeline is structured as a sequential workflow, with robust error handling for missing outputs, failed classifications, or malformed responses.

4 Evaluation and Results

We evaluated our pipeline on the provided test datasets, using accuracy, precision, recall, and F1-score as key metrics.

4.1 Sentiment Analysis Evaluation

- Accuracy: 0.650
- Precision: 0.635
- Recall: 0.664
- F1 Score: 0.634

These results indicate a moderate performance for sentiment analysis, with relatively balanced precision and recall. However, some misclassifications persist, particularly in distinguishing between mixed and strongly polarized sentiments.

4.2 Toxicity Detection Evaluation

- Accuracy: 0.560
- Precision: 0.292
- Recall: 0.467
- F1 Score: 0.359

Toxicity detection remains a key challenge, with a notably low precision (0.292), suggesting a high false positive rate. While recall (0.467) indicates some success in identifying toxic samples, the overall F1-score (0.359) highlights room for improvement. Refinements in model prompting, threshold tuning, or using an alternative toxicity classifier may enhance future performance.

5 Inter-Annotator Agreement Study

To evaluate the quality of detoxification, two annotators independently rated 15 detoxified samples on a scale from 1 to 10. The goal of this evaluation was to assess how effectively the detoxification model transformed toxic text into polite or neutral alternatives while maintaining the original meaning.

On average, Jacob rated the detoxifications at 8.2, while Tim provided a slightly lower average rating of 7.2. The overall mean score across both annotators was 7.7, indicating that most detoxifications were perceived as good to very good, but there were cases where improvements were needed.

We measured the agreement between annotators using several statistical metrics. The Pearson correlation coefficient (0.878) suggests a strong linear relationship between the two sets of ratings, meaning the annotators assigned similar values. Similarly, the Spearman correlation (0.809) and Kendall's Tau (0.705) demonstrate strong ordinal agreement, implying that the rankings of the samples were consistent between annotators.

However, Cohen's Kappa (0.136), which measures categorical agreement, was relatively low. This indicates that while both annotators generally agreed on the relative ranking of detoxifications, they sometimes placed ratings in different broad categories (e.g., one annotator marking a detoxification as "adequate" while the other marked it as "good").

In terms of absolute agreement, the annotators assigned identical scores for 4 out of 15 samples, while 73% of the samples (11 out of 15) had ratings within ±1 point of each other. This suggests that while there is strong alignment in overall detoxification assessment, finer distinctions in quality may be subjective.

A qualitative review of disagreements suggests that variation often arose when determining whether detoxified outputs retained subtle passive aggression or whether slight rewordings were sufficient to remove toxicity. Future improvements to annotation guidelines could provide clearer criteria for distinguishing between minor tone shifts and fully neutralized outputs.

These findings highlight the need for continued refinement in detoxification models, particularly in achieving consistency in neutralization without distorting meaning. Additional annotation rounds with more samples may help improve both inter-annotator agreement and our understanding of model behavior.

6 Challenges and Limitations

Our key challenges included:

- **Translation Model Selection**: While Toucan is optimized for African languages, some languages are still imperfectly mapped. Further refinements may be needed.
- **Classification Performance**: Toxicity detection performed below expectations (F1-score: 0.359). Future improvements may involve prompt tuning or model selection.
- **Detoxification Evaluation**: Manual annotation is time-consuming, and inter-rater agreement analysis is pending.

7 Conclusion and Future Work

Our upgraded pipeline successfully integrates language detection, translation, sentiment analysis, toxicity detection, and detoxification. While sentiment analysis performed well, toxicity detection and detoxification require further evaluation and refinement.

Future improvements include:

- Refining toxicity detection to improve classification accuracy.
- Expanding detoxification evaluation with inter-rater agreement analysis.
- Optimizing translation handling for under-represented languages.
- Experimenting with alternative detoxification models for better output quality.

Our current results provide a strong foundation for further research into automated text moderation and content filtering.