Marsan at PAN 2024 TextDetox: ToxiCleanse RL, Paving the Way for Toxicity-Free Online Discourse

Notebook for PAN at CLEF 2024

Maryam Najafi[1,2,*,†], Ehsan Tavan[2,†] and Simon Colreavy[1,†]

[1]Department of Computer Science and Information Systems, University of Limerick, Castletroy, V94 T9PX Limerick, Ireland
[2]NLP Department, Part AI Research Center, Tehran, Iran

Abstract

Addressing the pervasive issue of toxicity in online communication requires innovative solutions beyond mere identification and removal of harmful content. This paper presents our solution for the Multilingual Text Detoxification (TextDetox) shared task at PAN 2024. We, the MarSan_AI team, propose a novel approach termed ToxiCleanse RL, which employs Reinforcement Learning (RL), specifically Proximal Policy Optimization (PPO), in tandem with Large Language Models (LLMs), for detoxification through text style transfer (TST). Our method aims to automatically rewrite toxic messages while preserving their original meaning. By utilizing a toxicity-based reward model, we guide the RL fine-tuning process to effectively reduce the generation of toxic language. Empirical evaluation on English and Russian datasets demonstrates the superior performance of our approach compared to existing detoxification techniques, achieving a manual evaluation score of 0.89 (ranked 2nd) for English and 0.70 (ranked 7th) for Russian. These results underscore the potential of RL-based approaches in mitigating toxicity in online discourse, paving the way for safer and more inclusive digital environments.

Keywords

Large Language Models (LLMs), Reward Model, Supervised Fine-Tuning (SFT), Proximal Policy Optimization (PPO)

1. Introduction

Detecting toxicity and other harmful content, such as hate speech, insults, and threats, is a major focus in Natural Language Processing (NLP) research. However, merely identifying such content doesn't offer proactive solutions beyond removal. Today, social media platforms are also grappling with toxicity issues, often resorting to content blocking. We advocate for an approach where toxic messages are automatically re-written to maintain their meaningful content while removing toxicity, a process known as **detoxification**. This area has attracted considerable attention from NLP researchers and remains an active field of investigation [1, 2, 3, 4, 5, 6, 7].

Detoxification can be addressed through Text Style Transfer (TST). Style transfer involves the rewriting of text while altering one or several style attributes, such as authorship [8, 9, 10, 11] sentiment, or politeness [9, 12, 13]. However, it is important to note that changing these style attributes can sometimes significantly alter the meaning of a sentence. Despite this, many style transfer models aim to transform sentences into ones of a different style while retaining similarity on the same topic [14]. This presents a challenging yet intriguing task, as it requires striking a delicate balance between preserving original meaning and adjusting stylistic elements.

In the dynamic field of Artificial Intelligence (AI), the fusion of Large Language Models (LLMs) with Reinforcement Learning (RL) techniques shows great potential. Particularly, Proximal Policy Optimization (PPO), a subset of RL algorithms, has emerged as a powerful tool. This paper extensively explores the integration of LLMs with RL, along with Parameter-Efficient Fine-Tuning (PEFT), aiming to

CLEF 2024: Conference and Labs of the Evaluation Forum, September 9–12, 2024, Grenoble, France
*Corresponding author.
†These authors contributed equally.
maryam.najafi@ul.ie (M. Najafi); ehsan.tavan@partdp.ai (E. Tavan); simon.colreavy@ul.ie (S. Colreavy)
https://github.com/MaryNJ1995 (M. Najafi)
0000-0001-5025-2044 (M. Najafi); 0000-0003-1262-8172 (E. Tavan); 0000-0002-1795-6995 (S. Colreavy)
© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

CEUR-WS.org/Vol-3740/paper-269.pdf

generate completely neutral text samples while maintaining their original meaning. Within this study, we introduce the **ToxiCleanse RL** Approach, a strategy based on RL for mitigating text toxicity. To accurately assess the impact of detoxification methods on the provided text, we propose a reward model based on text similarity and toxicity levels, aiming to mitigate unintended biases related to various social identities during the RL phase. This reward model guides the policy to generate neutral samples that align more closely with ground truth samples. Empirical results demonstrate that utilizing RL for fine-tuning language models to optimize the non-toxicity reward effectively reduces the generation of toxic language, outperforming existing detoxification methods in the literature.

We, the MarSan_AI team, perform a large-scale evaluation of style transfer models on the Multilingual Text Detoxification shared task at PAN 2024 [15, 16], comparing our new models with baselines and state-of-the-art approaches. We release our code and data in our **GitHub**. Our contributions are structured as follows: Section 2 details the task and data description. Section 3 reviews related work. Section 4 introduces our model framework. Section 5.1 outlines the evaluation metrics, and finally, Section 5 presents the results.

2. Task & Data description

The Multilingual Text Detoxification (TextDetox) [15, 16] task for 2024 addressed the pressing need to combat toxicity in user-generated content on social media platforms. Unlike traditional approaches that often involved simply blocking or filtering toxic content, TextDetox encouraged a proactive approach by providing users with a neutralized version of their messages. With evaluation based on style transfer accuracy, content preservation, and fluency, participants were challenged to employ unsupervised and cross-lingual detoxification methods to tackle the diverse linguistic and cultural nuances of toxicity.

The TextDetox task provided datasets for English, Russian, and multilingual contexts. For each of the nine different languages, there were 1,000 parallel pairs available, split into development (400 pairs) and test (600 pairs) sets. Additionally, there were 19.7k English and 11.1k Russian data points available for the training phase. The datasets aimed to facilitate the development and evaluation of effective solutions for detoxifying toxic text across diverse linguistic contexts, contributing to a safer and more inclusive online environment. All submissions were managed through Codalab and tira.io [17].

3. Background

In [18], a groundbreaking method for detoxification leveraging parallel data was introduced. This innovative approach involved the creation of parallel datasets containing toxic sentences alongside their corresponding non-toxic paraphrases, both in English and Russian languages. Through meticulous crowdsourcing efforts, the authors curated over 10,000 non-toxic paraphrases for English toxic sentences, marking the inception of the first parallel datasets tailored explicitly for detoxification purposes. Furthermore, the study illustrated the process of distilling existing paraphrase datasets to derive toxic-neutral sentence pairs. By training detoxification models on these meticulously crafted datasets, the paper demonstrated substantial enhancements over prevailing unsupervised methods, underscoring the efficacy of harnessing parallel data in detoxification systems.

[4] Introduced two novel methods for removing toxicity from text. The first, ParaGeDi, employed style-guided language models and paraphrasing to retain content while eliminating toxicity. The second method, CondBERT, utilized BERT to replace toxic words with non-offensive alternatives. [1] Introduced pioneering methods for detoxifying Russian texts, marking a significant step in combating offensive language. Their innovative approaches, based on BERT and GPT-2 models, effectively transformed toxic content into neutral language. Through rigorous evaluation and comparison, the authors demonstrated the efficacy of their techniques, offering valuable contributions to content moderation in the Russian language. This study not only expanded the scope of TST tasks but also provided practical tools for fostering a safer online environment.

[19] introduced a method to extend text detoxification to multiple languages using parallel data. It is built upon existing techniques, showing the effectiveness of parallel corpora in improving text detoxification. The study also discussed the broader context of TST, highlighting the importance of parallel datasets in advancing research in this domain. By extending the ParaDetox pipeline to support multiple languages, including Russian, Ukrainian, and Spanish, the work aimed to facilitate safer communication in digital environments across linguistic boundaries. [20] follows an iterative process of leveraging human feedback to train summarization models. Initially, human preference data was collected by presenting evaluators with pairs of summaries and asking them to choose the better one. Then the reward model was trained to predict these human preferences, which was used as a reward function in a reinforcement learning setup, specifically employing the PPO algorithm. [21] enhances reinforcement learning from human feedback (RLHF) by introducing contrastive rewards. It involves two steps: offline sampling to obtain baseline responses and computing contrastive rewards based on these samples. These rewards enable self-improvement of the RL policy, penalizing uncertainty and improving robustness. Empirical testing demonstrates superior performance compared to standard RLHF, highlighting its effectiveness in aligning LLMs with human feedback. There are also other researches in this field [4, 22].

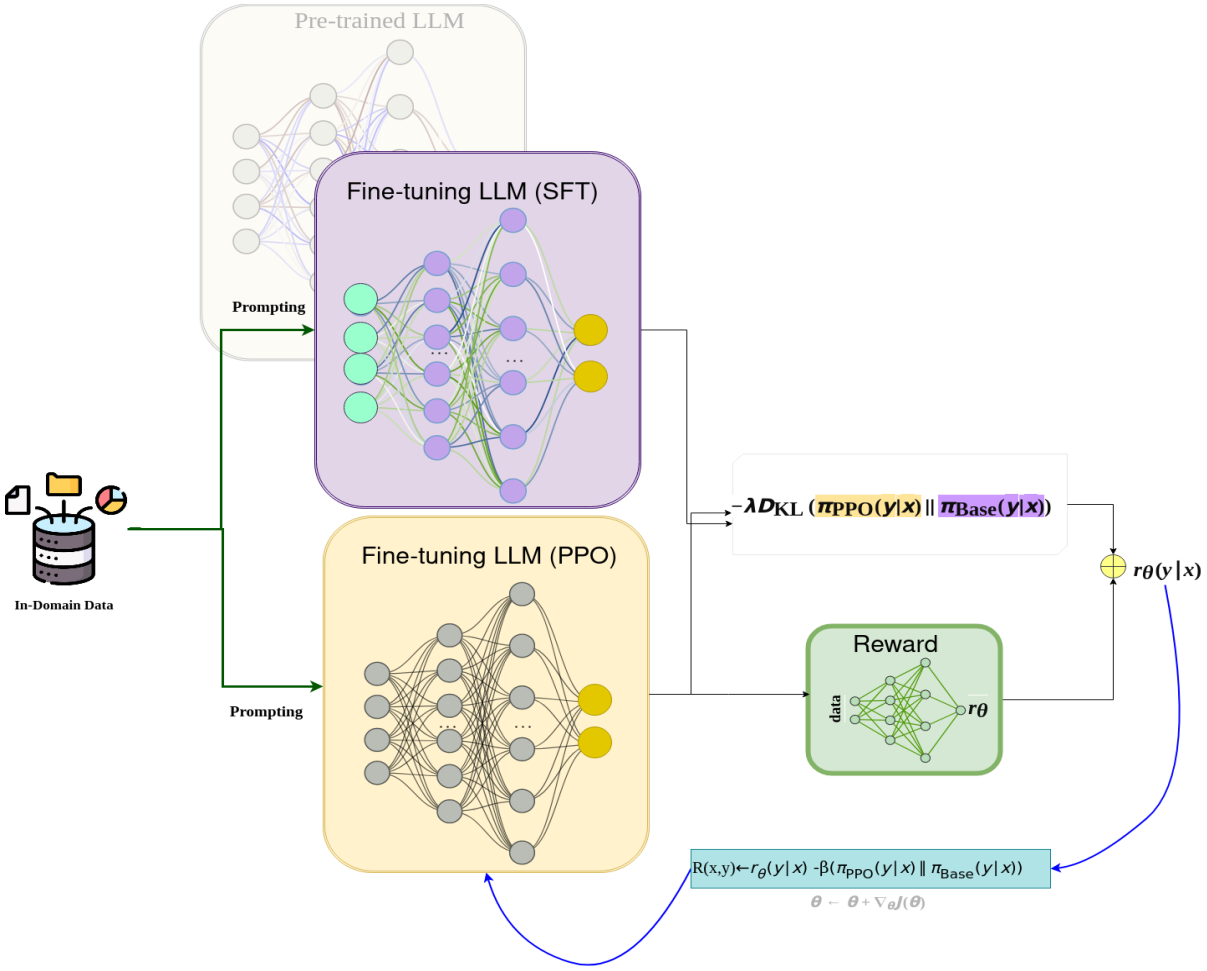

Figure 1: An overview of the proposed model.

4. System Overview

We provide an overview of our approach in Figure 1. This section delineates the fine-tuning process for generating less-toxic summaries using the Mistral LLM. Initially, we outline the base LLM architecture

and its parameters, providing a foundational understanding of the model and generating zero-shot samples. Subsequently, we detail the phases of Supervised Fine-Tuning (SFT) and the Proximal Policy Optimization (PPO) method.

4.1. Base LLM and Parameters

The Mistral 7B, introduced by Mistral AI, is a groundbreaking large language model available on the Hugging Face repository. It features advanced attention mechanisms like Sliding Window Attention (SWA) and Grouped-query Attention (GQA), optimizing both speed and memory usage. This design enables Mistral 7B to outperform larger models such as Llama 2 (13B) and Llama 1 (34B) on various benchmarks, making it versatile for commercial and research applications. Licensed under Apache 2.0, Mistral 7B is ideal for self-hosted AI solutions [23]. Although we started with Mistral as our initial model choice, we ultimately selected an upgraded version Mistral-T5-7B-v1 developed by Ignos for our final model. This advanced language model demonstrates exceptional performance in handling lengthy sequences, accommodating up to 32,768 tokens in context.

4.2. First Phase Fine-Tuning with SFT

We started with models pre-trained to autoregressively generate non-toxic samples. These pre-trained models served as 'zero-shot' baselines. Taking the prompts, the toxic sample, and the neutral corresponding sample as an example. In the next step, the Supervised Fine-tuning (SFT) model learns how to generate a neutral sample $(y) \sim \pi_{\text{SFT}}(y|x)$ based on the user's given toxic sample x. This process enables us to acquire a collection of baseline responses denoted as $\{y_{\text{base},i,j}\}_{j=1}^{k}$, where $y_{\text{base},i,j} \sim \pi_{\text{SFT}}(\cdot|x_i)$. These responses are then used as a comparison for measuring PPO model output in the evaluation phase. Hence, we fine-tuned these models via supervised learning on our competition datasets. These supervised models were used to generate initial neutral samples for collecting comparisons, to initialize our policy and reward models, and as baselines for evaluation.

4.3. Last Phase Fine-Tuning with PPO

In reinforcement learning with neural network function approximators, various approaches like deep Q-learning, vanilla policy gradient methods, and trust region or natural policy gradient methods have been explored. Each has strengths and weaknesses in scalability, data efficiency, and robustness across diverse problem domains. To address these challenges, Schulman et al. introduced the PPO algorithm [24].

PPO, a policy gradient method, combines the benefits of Trust Region Policy Optimization (TRPO) while simplifying implementation and enhancing sample efficiency. Policy gradient methods estimate the policy gradient and use it in a stochastic gradient ascent algorithm. The common gradient estimator is:

$$\hat{g} = \hat{\mathbb{E}}_t \left[\nabla_\theta \log \pi_\theta(a_t|s_t) \hat{A}_t \right]$$

where π_θ is a stochastic policy and $\hat{A}_t$ is an estimator of the advantage function at timestep t. The expectation $\hat{\mathbb{E}}_t[\cdot]$ denotes the empirical average over a finite batch of samples. However, performing multiple optimization steps on this loss using the same trajectory can lead to excessively large policy updates. Unlike conventional policy gradient methods that typically perform a single gradient update per data sample, PPO employs a surrogate objective function to enhance training stability by limiting the magnitude of policy updates and avoiding drastic changes. This is achieved by calculating a ratio indicating the difference between the current and old policies and then clipping this ratio within a specific range, $[1 - \epsilon, 1 + \epsilon]$. PPO ensures that policy updates remain conservative, promoting stable and reliable training progress.

Central to PPO is the clipped surrogate objective function, which stabilizes training by incorporating a constrained probability ratio between the current and old policies. This prevents overly large policy updates and ensures that gradient ascent steps encourage actions leading to higher rewards while

avoiding harmful actions. Determining the appropriate step size is critical: too small a step results in slow training, while too large a step introduces excessive variability. PPO addresses this by constraining policy updates within a small range using the clipped surrogate objective function, effectively avoiding destructive large-weight updates.

Let $r_t(\theta)$ denote the probability ratio:

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$$

This ratio represents the probability of taking action a_t in state s_t under the current policy, divided by the probability of taking the same action under the previous policy. Essentially, $r_t(\theta)$ measures the divergence between the old and current policies.

TRPO maximizes the surrogate objective:

$$L_{CPI}(\theta) = \hat{\mathbb{E}}_t \left[r_t(\theta)\hat{A}_t \right]$$

To avoid excessively large policy updates, PPO modifies this objective:

$$L_{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t \right) \right]$$

where ϵ is a hyperparameter. The clipping removes the incentive for r_t to move outside $[1-\epsilon, 1+\epsilon]$, ensuring stable updates.

By using the XLM-Roberta model's non-toxic class score as the reward in the PPO algorithm, we iteratively fine-tune the policy to generate non-toxic outputs. This approach ensures the generation of high-quality, detoxified content by leveraging the synergy between PPO updates and the toxicity-based reward model.

4.4. Reward Model

In our approach, we utilize an XLM-Roberta model fine-tuned for toxicity detection to compute rewards for the PPO algorithm. This model classifies input data into two classes: non-toxic (class index 0) and toxic (class index 1). The reward signal for the PPO model is derived from the output score of the XLM-Roberta model for the non-toxic class. Specifically, the score corresponding to class index 0 is used as the reward. A positive score indicates non-toxic content, while a negative score indicates toxic content. For instance, if the XLM-Roberta model assigns a score of +3.5 to a sample, it is considered non-toxic, and if it assigns a score of -2.8, the sample is considered toxic. This score is then used in the PPO update rule, guiding the model towards generating non-toxic content.

Using the reward model, we defined a policy to generate higher-quality outputs with reinforcement learning, maximizing the reward with the PPO algorithm. The policy was initialized using the supervised fine-tuned model. To encourage exploration and prevent the policy from deviating too much from the supervised model, we included a KL divergence term in the reward. This term serves as an entropy bonus, ensuring consistency with the training data.

The PPO value function used a Transformer with separate parameters from the policy to prevent pre-trained policy degradation during early training. The value function was initialized with reward model parameters. In our experiments, the reward model, policy, and value function were the same size.

The reward function $R(x, y)$ used in the PPO model to generate detoxified samples is defined as:

$$R(x, y) = r_\theta(x, y) - \beta \log \left(\frac{\pi_{\text{RL}_\phi}(y|x)}{\pi_{\text{SFT}}(y|x)} \right) \tag{1}$$

where $r_\theta(x, y)$ is the reward from the transformer model, β controls the penalty, $\pi_{\text{RL}_\phi}(y|x)$ is the PPO-trained policy probability, and $\pi_{\text{SFT}}(y|x)$ is the supervised fine-tuned policy probability. This logarithmic term penalizes deviations from the supervised model, encouraging optimal and detoxified outputs.

5. Results

In this section, we delve into the outcomes obtained from our analysis of the dev data. We explore the performance of various models and their efficacy in handling the detoxification task. Through comprehensive evaluation and analysis, we shed light on the strengths and weaknesses of each model, providing insights into their capabilities and potential areas for improvement.

5.1. Evaluation Metrics

Multiple attempts have been made to evaluate sentence style and toxicity, focusing on three key parameters of style transfer quality: text style, content preservation, and text fluency.[25] examine various detoxification models and explore the correlation between manual and automatic evaluation metrics, identifying metrics like ChrF and BertScore as potential proxies for human evaluation. For the competition, organizers provide automatic evaluation metrics have set.

- **Style Transfer Accuracy (STA):** This metric classifies the non-toxicity level in the generated paraphrase using a specifically fine-tuned xlm-roberta-large model for toxicity binary classification. Additionally, a base fine-tuned version of the classifier is provided for further experimentation.
- **Content Preservation (SIM):** This metric evaluates the content similarity between the original toxic sentence and the generated paraphrase by calculating the cosine similarity between LaBSe embeddings.
- **ChrF:** This metric estimates the text adequacy and its similarity to human-written detoxified references.
- **Joint:** combines the individual components of the automatic evaluation, is calculated as the mean of STA * SIM * FL per sample. This composite metric provides a unified measure of style transfer quality across the competition.

5.2. Quantitative Analysis of Dev Dataset Results

Table 1 showcases the performance of various language models (LLMs) in detoxifying toxic words in English dev datasets, evaluated across four metrics: STA, SIM, CHRF, and J. Mistral-T5-7B-v1 demonstrating the highest overall performance, with Mistral and Mistral-7B-Instruct-v0.1 also exhibiting strong results. Falcon-7b and its variants show moderate performance, while Lama2-13B and Lama2-7B display a considerable difference in performance. Zephyr-7b-beta and solar-10.7B-v1.0 perform well across all metrics. Similarly, Table 2 compares LLM performance for Russian dev datasets. Mistral-T5-7B-v1 leads, with competitive results from Mistral, falcon-7b-instruct, Lama2-13B, Lama2-7B, Zep, and solar. Zero-shot versions generally lag behind fine-tuned models, emphasizing the importance of tuning.

The Table 3 depicts the performance of English language models before and after data augmentation. Augmented versions show nuanced changes in metrics, with some models maintaining competitiveness while others exhibit slight variations. These results underscore the diverse effects of data augmentation on model performance, emphasizing the need for tailored strategies to optimize outcomes. Data augmentation can improve SIM (similarity) and CHRF (character n-gram F-score) metrics by diversifying the training data, thereby enhancing the model's ability to understand and generate text that aligns closely with the semantics and structure of the original text. Augmentation techniques allow the model to learn more robust representations of language and improve its performance in tasks requiring similarity and character-level understanding.

The Table 4 illustrates the comparison between SFT and PPO results for Mistral-T5-7B-v1 models in both English and Russian datasets. In English, PPO outperforms SFT across most metrics, with higher scores in STA, SIM, and J, indicating improved text understanding and generation capabilities. Similarly, in Russian, PPO shows enhancements in SIM, CHRF, and J scores compared to SFT. The improvements observed with PPO over SFT could be attributed to the reinforcement learning nature of PPO, which allows the model to iteratively adjust its parameters based on feedback from the environment, leading

Table 1
Performance of different LLMs on detoxification of toxic words for English DEV datasets

LLM Name	STA	SIM	CHRF	J
Mistral-7B-v0.1 Zero-shot	0.70	0.81	0.65	0.37
Mistral-7B-v0.1	0.77	0.85	0.73	0.491
Mistral-T5-7B-v1	**0.796**	**0.875**	0.745	**0.518**
Mistral-7B-Instruct-v0.1	0.77	0.85	0.71	0.487
falcon-7b Zero-shot	0.65	0.73	0.63	0.30
falcon-7b	0.68	0.85	0.72	0.417
falcon-7b-instruct	0.68	0.85	0.74	0.421
Lama2-7B	0.50	0.83	0.80	0.40
Lama2-13B	0.56	0.871	**0.81**	0.40
Lama3-8B Zero-shot	0.69	0.79	0.61	0.33
Lama3-8B	0.74	0.85	0.73	0.48
zephyr-7b-beta	0.76	0.85	0.70	0.49
solar-10.7B-v1.0	0.76	0.86	0.72	0.489

Table 2
Performance of different LLMs on detoxification of toxic words for Russian DEV datasets

LLM Name	STA	SIM	CHRF	J
Mistral-7B-v0.1 Zero-shot	0.65	0.80	0.63	0.32
Mistral-7B-v0.1	0.76	0.85	0.68	0.44
Mistral-T5-7B-v1	**0.80**	0.77	**0.76**	**0.47**
Mistral-7B-Instruct-v0.1	0.71	**0.87**	0.69	0.42
falcon-7b Zero-shot	0.61	0.79	0.61	0.28
falcon-7b	0.73	0.81	0.65	0.38
falcon-7b-instruct	0.69	0.86	**0.76**	0.45
Lama2-7B	0.65	0.84	0.70	0.39
Lama2-13B	0.67	0.86	0.71	0.40
Lama3-8B Zero-shot	0.69	0.79	0.61	0.33
Lama3-8B	0.67	0.86	0.71	0.40
zephyr-7b-beta	0.69	0.78	0.69	0.37
solar-10.7B-v1.0	0.67	0.85	0.73	0.41

Table 3
Results of data augmentations on English language models

LLM Name	STA	SIM	CHRF	J
Mistral-T5-7B-v1	0.796	0.875	0.745	0.518
Mistral-T5-7B-v1 (Augmented)	0.7	0.89	0.76	0.47
zephyr-7b-beta	0.78	0.85	0.72	0.493
zephyr-7b-beta (Augmented)	0.69	0.87	0.76	0.45
Lama3-8B - 2	0.74	0.85	0.73	0.48
Lama3-8B (Augmented)	0.67	0.87	0.75	0.43
Mistral-7B-v0.1	0.77	0.85	0.73	0.491
Mistral-7B-v0.1 (Augmented)	0.73	0.87	0.75	0.476

to more efficient learning and better adaptation to the given task and dataset. Additionally, PPO's ability to explore and exploit the training data more effectively might contribute to its superior performance over SFT. Our approach was evaluated using both automatic and manual evaluation metrics provided by the PAN organizers.

Table 4
Comparison of SFT and PPO results for Mistral-T5-7B-v1 models

	STA	SIM	CHRF	J
SFT/EN	0.796	0.875	0.745	0.518
PPO/EN	**0.808**	**0.89**	**0.76**	**0.541**
SFT/RU	0.8	0.77	0.76	0.47
PPO/RU	**0.8**	**0.79**	**0.79**	**0.499**

5.3. Automatic Evaluation Test Results

The automatic evaluation was based on the **J** evaluation metric. The results of the test data from the PAN organizers' leaderboard are presented in 5:

Table 5
Results of Test data(Automatic Evaluation)

Model/Language	J Score
PPO/EN	0.504
PPO/RU	0.508

The scores obtained from the automatic evaluation closely resemble our own internal evaluations.

5.4. Manual Evaluation Test Results

Manual evaluation was conducted through crowdsourcing on a random subsample of 100 texts per language. Our team achieved second place in the leaderboard for English data detoxification. The manual evaluation results on the test data from the PAN organizers' leaderboard are presented in 6.

Table 6
Results of Test data(Manual Evaluation)

Model/Language	Manual Evaluation Score (Rank)
PPO/EN	**0.89** (2)
PPO/RU	**0.70** (7)

Our approach performed competitively, especially in the field of English data cleansing. The final results highlighted the effectiveness of our RL-based method, which achieved second place in an English dataset. It's noteworthy that our objective was to assess model quality by fine-tuning large language models on substantial datasets. We focused exclusively on Russian and English due to the availability of high-quality open-source datasets for these languages.

6. Conclusion & Future Work

In summary, our research presents a holistic strategy for addressing toxicity in text through the application of RL, specifically PPO, in conjunction with LLMs. Our ToxiCleanse RL Approach utilizes RL fine-tuning to generate neutral text outputs while preserving their original meaning. Using a toxicity-based reward model, we successfully mitigate the generation of toxic language, surpassing existing detoxification methods and even outperforming Supervised Fine-tuned LLMs. Our findings underscore the effectiveness of RL-based rewards in elevating the quality of generated content.

Looking ahead, future endeavors could involve refining LM/LLM-based rewards through manual fine-tuning to enhance reward model accuracy further. Moreover, developing a similarity-based reward that measures detoxification while penalizing deviations between original and generated samples could prove instrumental in maintaining text integrity. These initiatives are poised to propel advancements in text detoxification, fostering safer and more inclusive online environments.

References

[1] D. Dementieva, D. Moskovskiy, V. Logacheva, D. Dale, O. Kozlova, N. Semenov, A. Panchenko, Methods for detoxification of texts for the russian language, Multimodal Technologies and Interaction 5 (2021) 54.

[2] D. Moskovskiy, D. Dementieva, A. Panchenko, Exploring cross-lingual text detoxification with large multilingual language models., in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, 2022, pp. 346–354.

[3] A. Wang, M. Sudhakar, Y. Ji, Simple text detoxification by identifying a linear toxic subspace in language model embeddings, arXiv preprint arXiv:2112.08346 (2021).

[4] D. Dale, A. Voronov, D. Dementieva, V. Logacheva, O. Kozlova, N. Semenov, A. Panchenko, Text detoxification using large pre-trained neural models, arXiv preprint arXiv:2109.08914 (2021).

[5] S. Mukherjee, A. Bansal, A. K. Ojha, J. P. McCrae, O. Dušek, Text detoxification as style transfer in english and hindi, arXiv preprint arXiv:2402.07767 (2024).

[6] J. Qian, Text Detoxification in Natural Language Processing, University of California, Santa Barbara, 2022.

[7] S. Jain, G. Kaushik, P. Prabhu, A. Godbole, Detox: Nlp based classification and euphemistic text substitution for toxic comments, in: 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), IEEE, 2021, pp. 1–5.

[8] M. Najafi, E. Tavan, Text-to-text transformer in authorship verification via stylistic and semantical analysis., in: CLEF (Working Notes), 2022, pp. 2607–2616.

[9] E. Tavan, M. Najafi, R. Moradi, Identifying ironic content spreaders on twitter using psychometrics, contextual and ironic features with gradient boosting classifier., in: CLEF (Working Notes), 2022, pp. 2687–2697.

[10] M. Najafi, S. Sadidpur, Paa: Persian author attribution using dense and recursive connection (2024).

[11] H. B. Giglou, T. Rahgooy, M. Rahgouy, J. Razmara, Uot-uwf-partai at semeval-2021 task 5: Self attention based bi-gru with multi-embedding representation for toxicity highlighter, arXiv preprint arXiv:2104.13164 (2021).

[12] M. Najafi, E. Tavan, Marsan at semeval-2022 task 6: isarcasm detection via t5 and sequence learners, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), 2022, pp. 978–986.

[13] E. Tavan, M. Najafi, Marsan at semeval-2023 task 10: Can adversarial training with help of a graph convolutional network detect explainable sexism?, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 1011–1020.

[14] D. Dale, A. Voronov, D. Dementieva, V. Logacheva, O. Kozlova, N. Semenov, A. Panchenko, Text detoxification using large pre-trained neural models (2021), arXiv preprint arXiv:2109.08914 (????).

[15] J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, P. Rosso, A. Smirnova, E. Stamatatos, B. Stein, M. Taulé, D. Ustalov, M. Wiegmann, E. Zangerle, Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.

[16] D. Dementieva, D. Moskovskiy, N. Babakov, A. A. Ayele, N. Rizwan, F. Schneider, X. Wang, S. M. Yimam, D. Ustalov, E. Stakovskii, A. Smirnova, A. Elnagar, A. Mukherjee, A. Panchenko, Overview of the multilingual text detoxification task at pan 2024, in: CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[17] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in

Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. URL: https://link.springer.com/chapter/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6_20.

[18] V. Logacheva, D. Dementieva, S. Ustyantsev, D. Moskovskiy, D. Dale, I. Krotova, N. Semenov, A. Panchenko, Paradetox: Detoxification with parallel data, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 6804–6818.

[19] D. Dementieva, N. Babakov, A. Panchenko, Multiparadetox: Extending text detoxification with parallel data to new languages, arXiv preprint arXiv:2404.02037 (2024).

[20] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, P. F. Christiano, Learning to summarize with human feedback, Advances in Neural Information Processing Systems 33 (2020) 3008–3021.

[21] W. Shen, X. Zhang, Y. Yao, R. Zheng, H. Guo, Y. Liu, Improving reinforcement learning from human feedback using contrastive rewards, arXiv preprint arXiv:2403.07708 (2024).

[22] D. Dementieva, D. Moskovskiy, D. Dale, A. Panchenko, Exploring methods for cross-lingual text style transfer: The case of text detoxification, arXiv preprint arXiv:2311.13937 (2023).

[23] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).

[24] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347 (2017).

[25] D. Dementieva, V. Logacheva, I. Nikishina, A. Fenogenova, D. Dale, I. Krotova, N. Semenov, T. Shavrina, A. Panchenko, Russe-2022: Findings of the first russian detoxification shared task based on parallel corpora (????).