PAN 2024 Multilingual TextDetox: Exploring Cross-lingual Transfer Using Large Language Models*

Notebook for PAN at CLEF 2024

Vitaly Protasov[1]

[1] *Artificial Intelligence Research Institute (AIRI), Moscow, Russia*

Abstract

Text detoxification is a text-to-text generation task that relies on available data for experiments. In recent years, this task has primarily focused on well-resourced languages while neglecting lower-resource languages. This work explores various approaches to building a multilingual solution for different languages, with an emphasis on 9 languages in the Multilingual Text Detoxification Task at PAN 2024. Throughout the experiments, we consider not only different model types but also employ fine-tuning on various combinations of datasets. As a result, we achieve third place in human evaluation and show promising progress towards developing a multilingual solution for the text detoxification task using large language models such as mT0 and XGLM. We also observe that fine-tuning on combinations of relatively similar languages is a promising direction—especially when real data for some languages is lacking.

Keywords

PAN 2024, Multilingual Text Detoxification (TextDetox) 2024, cross-lingual transfer, large language models

1. Introduction

Text detoxification is the process of rewriting a given text to remove or rephrase toxic or rude elements, making it more respectful and appropriate for a wider audience. This task has gained significant attention due to the growing concern for creating a safer and more inclusive online environment [1]. The text detoxification task presents several challenges. One primary issue is establishing clear and formalized criteria for defining inappropriate content. Another challenge lies in deciding the appropriate action for detected inappropriate parts of a sentence—whether they should be deleted, rewritten, or preserved—and whether the original meaning should be revised. While annotation criteria can help address these concerns, ensuring consistent understanding across different research studies and proposed datasets is crucial for making detoxification methods more deterministic and stable.

Another significant challenge here is the application of detoxification methods across various languages [2]. Current research predominantly focuses on high-resource languages such as English and Russian, while languages with fewer resources and data remain underrepresented. To address this gap, the PAN at CLEF 2024 has introduced a Multilingual Text Detoxification task [3, 4], offering data for 9 languages to support and advance research and development in this critical area.

In this study, we focus on developing a multilingual solution for the text detoxification task. We explore various methodologies, including training of encoder-decoder models such as mBART[5] and mT0 [6], as well as the training of decoder-only Large Language Model (LLM) XGLM[7]. Additionally, we conduct experiments on cross-lingual transfer by training models with different language combinations to achieve the highest performance on the test evaluation set. The combination of predictions from such models as mT0 and XGLM proved to be our best solution, securing fourth place in the test stage based on automatic evaluation and third place according to manual human evaluation.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

* You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

* Corresponding author.

† These authors contributed equally.

vitasprotas@gmail.com (V. Protasov)

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

2. Related work

In earlier works, unsupervised methods like CondBERT [8] demonstrated effectiveness in the text detoxification task by identifying and rephrasing toxic parts of the text. However, these methods were eventually surpassed by encoder-decoder approaches [9, 2]. Consequently, newer methods emerged, treating detoxification similar to machine translation task, where toxic text is the input and the detoxified version is the output. This approach has led to the inclusion of a growing number of languages, fostering solutions for languages previously unaddressed. For example, [10] organized a competition centered on detoxifying Russian text, highlighting various methods, including decoder-only models. Additionally, [11] introduced a dataset for English and proposed new detoxification methods. Following this, [12] investigated strategies for transferring knowledge to new languages using translation models as an intermediary step.

Our work aims to address the underrepresentation of methods for different languages by considering a new dataset for nine languages.

3. Experimental setup

3.1. Models

Most recent high-performing text detoxification approaches rely on encoder-decoder models. We decided to start with models of this architecture and include decoder-only LLMs, which have shown impressive results in natural language processing (NLP)[13]. Our experiments consider encoder-decoder models such as mBART and mT0, while for the decoder-only approach, we focus on XGLM. Previous studies have highlighted the effectiveness of models like mBART and mT5[14], but we decide to also consider mT0 due to its multitask fine-tuning capability, which could be advantageous for text detoxification. Additionally, in our selection of decoder-only model, we choose XGLM because of its multilingual nature and good reported performance results. It's worth noting that XGLM was not trained on Amharic data; thus, no results are reported in this language here. Regarding the mT0, we aim to consider different model sizes such as base[1], large[2], and xl[3]. Also, we consider XGLM (7.5B)[4] only.

3.2. Fine-tuning pipeline

Along with the dataset provided for the development[5], we utilize both English[6] and Russian[7] ParaDetox [11] datasets during the experiments. Our objective is to utilize these monolingual datasets to tailor models for text detoxification before fine-tuning on the provided multilingual dataset. As previously stated in Section 1, potential discrepancies exist in the collected datasets. Therefore, we do not consider merging such datasets for fine-tuning and decide to utilize them separately.

Though we consider both encoder-decoder and decoder-based models, the fine-tuning process in general is not necessarily different. We examine two types of datasets during fine-tuning: utilizing English and Russian ParaDetox datasets as well as the MultilingualParaDetox dataset for nine languages. With the first ones, we explore whether initial fine-tuning on English and Russian languages improves the convergence of multilingual models during final fine-tuning on the MPD dataset. Finally, we conduct fine-tuning on a multilingual dataset using different language combinations: (i) monolingual fine-tuning on each language independently; (ii) multilingual fine-tuning using all available languages; (iii) fine-tuning on different combinations of languages. For the third approach, our hypothesis suggests that closely related languages significantly impact fine-tuning success more than a monolingual approach.

[1] bigscience/mt0-base
[2] bigscience/mt0-large
[3] bigscience/mt0-xl
[4] facebook/xglm-7.5B
[5] textdetox/multilingual_paradetox
[6] https://huggingface.co/datasets/s-nlp/paradetox
[7] https://huggingface.co/datasets/s-nlp/ru_paradetox

Also, since the MPD lacks annotated data—with only 400 examples available for each language—various combinations of similar languages should be considered to augment training data and may lead to improved performance on holdout sets.

Due to the constraints of computational resources for LLMs like mT0-xl and XGLM, we choose not to fine-tune all their weights. Instead, we investigate the potential of utilizing Low Rank Adaptation (LoRA) [15] to facilitate the training of these large models.

4. Results

During the experiments, prompts are not used, and we rely solely on model convergence without providing additional instructions during fine-tuning and evaluation. Additionally, early stopping [16] is employed. All reported results are based on metric values measured automatically on the test set.

We iteratively test our hypotheses without considering all potential scenarios within each experiment iteration, so we do not present results for every experimental setup.

4.1. Multilingual fine-tuning

First, we aim to investigate the multilingual performance of various models and consider fine-tuning them on different combinations of datasets, including PD and MPD. Table 1 presents the results of these experiments. Notably, utilizing the PD dataset can enhance target performance in most languages after fine-tuning mBART and mT0-large.

For other models, we decided not to explore different combinations of datasets due to the large size and complexity of experiments. Thus, for them we only included results for PD and MPD together. As we can see, mT0-large achieves the maximum absolute values in target performance in five languages; mT0-xl+LoRA is the best for one language only, and XGLM+LoRA performs the best in three languages.

Table 1
Test results of different multilingual models based on the datasets used during fine-tuning.

Model	Dataset	en	es	de	zh	ar	hi	uk	ru	am
mBART	MPD	0.36	**0.28**	**0.31**	0.05	0.19	0.11	0.24	0.27	0.08
mBART	PD+MPD	**0.38**	0.26	0.28	**0.07**	**0.24**	**0.14**	**0.25**	**0.34**	**0.11**
mT0-base	MPD	0.35	0.25	0.28	**0.07**	0.31	0.12	0.31	0.30	0.09
mT0-base	PD+MPD	**0.38**	**0.29**	**0.32**	0.06	**0.36**	**0.13**	**0.34**	**0.32**	**0.11**
mT0-large	PD+MPD	**0.48**	0.39	0.42	**0.11**	0.43	0.2	**0.48**	**0.46**	**0.17**
mT0-xl+LoRA	PD+MPD	0.41	0.41	**0.45**	0.10	0.47	0.16	0.43	0.41	0.15
XGLM+LoRA	PD+MPD	0.44	**0.466**	0.43	0.08	**0.505**	**0.303**	0.42	0.40	-

4.2. Fine-tuning across different combination of languages

In Section 4.1, we found that mT0-large and XGLM+LoRA showed the most promising performance. However, conducting experiments with LLMs like XGLM requires significant computational resources and time. Since our focus in this section is to explore fine-tuning across various language combinations, we choose to conduct experiments here with mT0-large only due to the trade-off between its size and multilingual performance.

In Table 2, we present the results obtained from fine-tuning using different combinations of languages. Specifically, we observe that fine-tuning the model in Russian or Ukrainian separately yields poorer performance than fine-tuning their combination. Similar patterns were observed in experiments with Hindi and Amharic, where training on their combination resulted in the best performance. However, when it comes to German, English, Spanish, and Arabic, fine-tuning using their combination shows worse results than fine-tuning them separately. Nevertheless, English and Spanish still exhibit similar improvement patterns when combined.

As a result, we can approve our hypothesis that closely related languages can serve to improve results while training on their combination, enriching the training dataset.

Table 2
Test results for the mT0-large based on the combination of languages used during fine-tuning.

Dataset	en	es	de	zh	ar	hi	uk	ru	am
PD RU + MPD RU	-	-	-	-	-	-	-	0.456	-
PD RU + MPD UK	-	-	-	-	-	-	0.543	-	-
PD RU + MPD RU+UK	-	-	-	-	-	-	**0.583**	**0.525**	-
MPD HI	-	-	-	-	-	0.245	-	-	-
MPD AM	-	-	-	-	-	-	-	-	0.285
MPD HI+AM	-	-	-	-	-	**0.274**	-	-	**0.298**
PD EN + MPD EN	**0.525**	-	-	-	-	-	-	-	-
MPD DE	-	-	**0.502**	-	-	-	-	-	-
PD EN + MPD EN+DE	0.471	-	0.481	-	-	-	-	-	-
MPD ES	-	0.35	-	-	-	-	-	-	-
PD EN + MPD ES+EN	0.498	**0.39**	-	-	-	-	-	-	-
MPD AR	-	-	-	-	0.502	-	-	-	-
MPD ES+AR	-	0.37	-	-	0.491	-	-	-	-

4.3. Excluding toxic lexicon from combined results

Based on the reported results, our final submission combines the top-performing outputs from the mt0-large model for English, German, Ukrainian, Russian, and Amharic with those from XGLM+LoRA for Spanish, Arabic, and Hindi. Although we could not exceed the performance of the *delete baseline* for Chinese, we have replicated its results and included them in our final submission.

Afterward, we chose to preprocess these combined results by excluding words from a multilingual toxic lexicon dataset[8] provided in the competition. Table 3 illustrates a comparison of the results before and after excluding toxic lexicon words at this stage. As we can see, the removal of such words positively impacts almost all languages, though it did not affect the results for Chinese and German.

Table 3
Test results of the submission with the best combined outcomes before and after excluding toxic lexicon words.

	en	es	de	zh	ar	hi	uk	ru	am
Before	0.525	0.466	**0.502**	**0.175**	0.505	0.303	0.583	0.525	0.298
After	**0.531**	**0.472**	**0.502**	**0.175**	**0.523**	**0.320**	**0.629**	**0.542**	**0.311**

4.4. Manual evaluation results

As mentioned earlier, our top submission secured fourth place in the automatic test evaluation, yet it reached third place in the manual evaluation through human annotation (refer to Table 4). Notably, according to human evaluation, our results for such languages as Spanish, Hindi, and Arabic are the top ones, indicating that decoder-only LLMs are more effective at handling the text detoxification task and generating more human-like text.

5. Conclusion

This study explored fine-tuning various models with different architectures for the task of text detoxification. Our experiments also investigated the use of varied combinations of datasets and languages

[8] textdetox/multilingual_toxic_lexicon

during fine-tuning. By combining different approaches, we achieved fourth place in test evaluation and third place in human evaluation. In this work, we particularly demonstrated that cross-lingual transfer between languages is a promising approach, improving languages such as Ukrainian and Amharic by transferring knowledge from closely related languages such as Russian and Indian respectively. We also showed that training decoder-only LLMs can be a promising direction, yielding the best results according to human evaluation, which totally aligns with the latest advancements in the NLP sphere.

Table 4
The leaderboard with first 5 participants ranked by the target average metric. Top-1 results are highlighted by bold and underline; Top-3 participants are highlighted by bold.

Participant	average	en	es	de	zh	ar	hi	uk	ru	am
Human References	0.85	0.88	0.79	0.71	0.93	0.82	0.97	0.90	0.80	0.85
SomethingAwful	0.77	0.86	**0.83**	**0.89**	0.53	0.74	**0.86**	**0.69**	**0.84**	**0.71**
adugeen	0.74	0.83	0.73	0.70	0.60	**0.82**	0.68	**0.84**	**0.76**	**0.71**
VitalyProtasov	0.72	0.69	**0.81**	0.77	0.49	**0.79**	**0.87**	0.67	0.73	0.68
nikita.sushko	0.71	0.70	0.62	**0.79**	0.47	**0.89**	**0.84**	0.67	0.74	0.68
erehulka	0.71	0.88	0.71	**0.85**	**0.68**	0.78	0.52	0.63	0.65	0.69

References

[1] G. Floto, M. M. T. pour, P. Farinneya, Z. Tang, A. Pesaranghader, M. Bharadwaj, S. Sanner, Diffudetox: A mixed diffusion model for text detoxification, ArXiv abs/2306.08505 (2023). URL: https://api.semanticscholar.org/CorpusID:259164399.

[2] D. Moskovskiy, D. Dementieva, A. Panchenko, Exploring cross-lingual text detoxification with large multilingual language models., ArXiv abs/2206.02252 (2022). URL: https://api.semanticscholar.org/CorpusID:249394890.

[3] 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: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.

[4] 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.

[5] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer, Multilingual denoising pre-training for neural machine translation, Transactions of the Association for Computational Linguistics 8 (2020) 726–742. URL: https://api.semanticscholar.org/CorpusID:210861178.

[6] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. L. Scao, M. S. Bari, S. Shen, Z.-X. Yong, H. Schoelkopf, X. Tang, D. R. Radev, A. F. Aji, K. Almubarak, S. Albanie, Z. Alyafeai, A. Webson, E. Raff, C. Raffel, Crosslingual generalization through multitask finetuning, in: Annual Meeting of the Association for Computational Linguistics, 2023. URL: https://api.semanticscholar.org/CorpusID:253264914.

[7] X. V. Lin, T. Mihaylov, M. Artetxe, T. Wang, S. Chen, D. Simig, M. Ott, N. Goyal, S. Bhosale, J. Du, R. Pasunuru, S. Shleifer, P. S. Koura, V. Chaudhary, B. O'Horo, J. Wang, L. Zettlemoyer, Z. Kozareva,

M. T. Diab, V. Stoyanov, X. Li, Few-shot learning with multilingual language models, ArXiv abs/2112.10668 (2021). URL: https://api.semanticscholar.org/CorpusID:260651613.

[8] D. Dale, A. Voronov, D. Dementieva, V. Logacheva, O. Kozlova, N. Semenov, A. Panchenko, Text detoxification using large pre-trained neural models, ArXiv abs/2109.08914 (2021). URL: https://api.semanticscholar.org/CorpusID:237572304.

[9] L. Laugier, J. Pavlopoulos, J. S. Sorensen, L. Dixon, Civil rephrases of toxic texts with self-supervised transformers, ArXiv abs/2102.05456 (2021). URL: https://api.semanticscholar.org/CorpusID:231861515.

[10] V. Logacheva, D. Dementieva, I. Krotova, A. Fenogenova, I. Nikishina, T. Shavrina, A. Panchenko, A study on manual and automatic evaluation for text style transfer: The case of detoxification, Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval) (2022). URL: https://api.semanticscholar.org/CorpusID:248780050.

[11] V. Logacheva, D. Dementieva, S. Ustyantsev, D. Moskovskiy, D. Dale, I. V. Krotova, N. Semenov, A. Panchenko, Paradetox: Detoxification with parallel data, in: Annual Meeting of the Association for Computational Linguistics, 2022. URL: https://api.semanticscholar.org/CorpusID:248780527.

[12] D. Dementieva, D. Moskovskiy, D. Dale, A. Panchenko, Exploring methods for cross-lingual text style transfer: The case of text detoxification, in: International Joint Conference on Natural Language Processing, 2023. URL: https://api.semanticscholar.org/CorpusID:265445167.

[13] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Barnes, A. S. Mian, A comprehensive overview of large language models, ArXiv abs/2307.06435 (2023). URL: https://api.semanticscholar.org/CorpusID:259847443.

[14] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, in: North American Chapter of the Association for Computational Linguistics, 2020. URL: https://api.semanticscholar.org/CorpusID:225040574.

[15] J. E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, Lora: Low-rank adaptation of large language models, ArXiv abs/2106.09685 (2021). URL: https://api.semanticscholar.org/CorpusID:235458009.

[16] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, N. A. Smith, Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping, ArXiv abs/2002.06305 (2020). URL: https://api.semanticscholar.org/CorpusID:211132951.