Multilingual Text Detoxification Using Google Cloud Translation and Post-Processing

Notebook for PAN at CLEF 2024

Zhongyu Luo[1], Man Luo[2] and Aiguo Wang[1,*]

[1]*Foshan University, Foshan, Guangdong, China*
[2]*Guangzhou City University of Technology, Guangzhou, Guangdong, China*

Abstract

The task of text detoxification aims to re-write toxic text into non-toxic text. Though existing methods have achieved impressive detoxification performance in monolingual settings, multilingual text detoxification remains challenging due to the complexity of natural languages and the lack of sufficient data for training accurate models for minor languages. In this study, we propose a cross-lingual text detoxification model, named GCTP, utilizing Google Cloud Translation and post-processing for the PAN@CLEF 2024 multilingual text detoxification task. Specifically, GCTP first translates minor language text into English for detoxification with a pretrained English model, and then translates it back to remove toxic keywords with predefined dictionaries of the original language. Extensive comparative experiments on competition datasets show that GCTP achieves the highest $\mathcal{J}$ score for Amharic text and ranks the 5th place for Chinese text, demonstrating the effectiveness of GCTP.

Keywords

Multilingual Text Detoxification, Text Detoxification, Google Cloud Translation, Minor Language

1. Introduction

With the widespread use and development of social networks, forums, and online communication platforms, the prevalence of toxic language has significantly increased, necessitating effective methods to mitigate offensive content. Text detoxification involves modifying certain attributes of the text while preserving its semantic content. For the PAN@CLEF 2024 multilingual text detoxification task, the goal is to rewrite toxic text into non-toxic text while maintaining the main content as much as possible [1, 2]. Though some methods have achieved promising results in monolingual settings [3], the semantic differences between languages make cross-lingual detoxification a challenging problem [4, 5]. Figure 1 lists examples of detoxification.

Existing detoxification methods mainly use predefined rules to delete toxic words or phrases [6]. However, these approaches often fail to consider the overall meaning of a sentence and produce unnatural outputs [7, 8]. Accordingly, researchers have explored deep learning-based text detoxification models to address these issues. For example, Pletenev et al. proposed autoregressive and non-autoregressive models to detoxify Russian text, treating detoxification as a specific case of text style transfer. Their approach uses an automatic post-editing algorithm to correct toxic text [9]. Totmina et al. used the pre-trained GPT3 to detoxify Russian text, highlighting the importance of filtering training data and fine-tuning models for specific languages [10]. Besides the exploration of English and Russian text [11, 12], Dementieva et al. explored the potential of large multilingual models (e.g., mBART and mT5) in cross-lingual text style transfer [13]. Their work emphasizes the multilingual and cross-lingual detoxification challenges and limitations. Motivated by previous studies, we propose a cross-lingual text detoxification model, named GCTP, to tackle the PAN@CLEF 2024 multilingual text detoxification task.

The contributions of our work are as follows:

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

[*] Corresponding author.

luozhongyu2799@163.com (Z. Luo); joeyluo9527@163.com (M. Luo); wangaiguo2546@163.com (A. Wang)

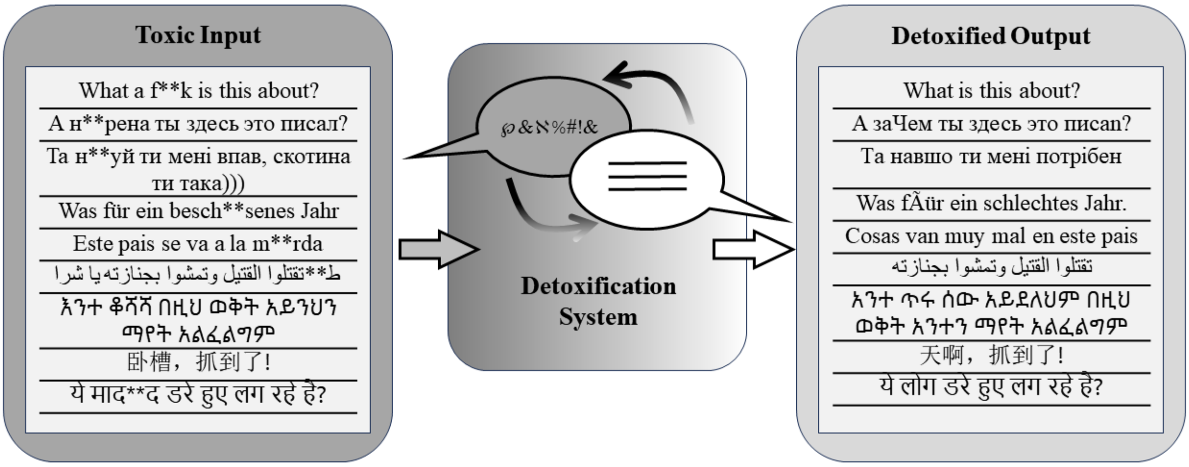

Toxic Input
What a f**k is this about?
А н**рена ты здесь это писал?
Та н**уй ти мені впав, скотина ти така)))
Was für ein besch**senes Jahr
Este pais se va a la m**rda
ط**تقتلوا القتيل وتمشوا بجنازته يا شرا
እንተ ቆሻሻ በዚህ ወቅት አይንህን ማየት አልፈልግም
卧槽，抓到了!
ये माद**द डरे हुए लग रहे है?
℘&ℵ%#!&
Detoxification System
Detoxified Output
What is this about?
А заЧем ты здесь это писan?
Та навшо ти мені потрібен
Was fÃür ein schlechtes Jahr.
Cosas van muy mal en este pais
تقتلوا القتيل وتمشوا بجنازته
አንተ ጥሩ ሰው አይደለህም በዚህ ወቅት አንተን ማየት አልፈልግም
天啊，抓到了!
ये लोग डरे हुए लग रहे है?

Figure 1: Examples of Detoxification Results

(1) Motivated by the transfer learning paradigm, we use Google Cloud Translation API to translate minor language text into English text and use a pretrained English model (i.e., Bart-base-detox in the task) to detoxify text at the semantic level. The processed text is then translated into the minor language, followed by the post-processing that deletes toxic text with a dictionary.
(2) Comparative experiments are conducted on the competition datasets and compared with three baseline detoxification methods. Experimental results achieves the highest $\mathcal{J}$ score for Amharic text in the task.

The structure of this paper is as follows. Section 2 details the proposed model text detoxification. Section 3 describes the datasets and preprocessing steps , outlines the experimental setup and baseline methods, and gives results and detoxification examples. Section 4 concludes this study.

2. Methodology

To detect toxic text at the semantic level, a large language model is preferred and hence in this study, we propose a cross-lingual text detoxification model. Specifically, if a text detoxification model is available for a language (e.g., English and Russian in the contest), we first use the detoxification model to rewrite toxic text into non-toxic text and then delete toxic text with predefined toxic keywords. In this study, we use the Bart-base-detox model and the ruT5-base-detox model to detoxify English and Russian text, respectively.

For minor languages (e.g., Spanish, German, Chinese, Arabic, Hindi, Ukrainian, and Amharic in the contest), since there is no corresponding detoxification model, GCTP leverages the translation and detoxification capabilities of existing tools to detoxify multilingual text. Specifically, GCTP mainly consists of four steps: Translation, Detoxification, Backtranslation, and Post-processing, as shown in Figure 2.

(1) **Translation**. Translate the minor language text (i.e., target domain) into English (i.e., source domain) to utilize the powerful detoxification models available for the English language. Google Cloud Translation API is used to convert the minor language text into English.
(2) **Detoxification**. Detoxify the translated English text using a pretrained model specialized in removing toxic content. The Bart-base-detox model, a pretrained English detoxification model, is used to handle the translated English text. Bart-base-detox model can neutralize toxic language while preserving the semantic meaning of the original text. Its detoxification process involves the following steps:
    (a) **input processing**: the translated English text is tokenized and fed into the Bart-base-detox model.

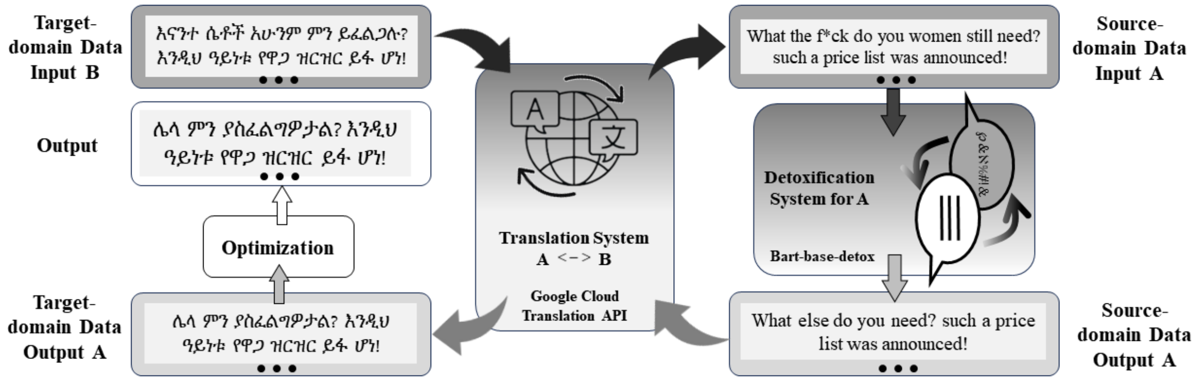

Target-domain Data Input B
እናንተ ሴቶች አሁንም ምን ይፈልጋሉ?
እንዲህ ዓይነቱ የዋጋ ዝርዝር ይፋ ሆነ!
Output
ሌላ ምን ያስፈልግዎታል? እንዲህ
ዓይነቱ የዋጋ ዝርዝር ይፋ ሆነ!
Optimization
Target-domain Data Output A
ሌላ ምን ያስፈልግዎታል? እንዲህ
ዓይነቱ የዋጋ ዝርዝር ይፋ ሆነ!
Translation System
A <-> B
Google Cloud Translation API
What the f*ck do you women still need?
such a price list was announced!
Source-domain Data Input A
Detoxification System for A
Bart-base-detox
&!#%N&&
What else do you need? such a price
list was announced!
Source-domain Data Output A

Figure 2: Transfer Learning Enabled Cross-Lingual Text Detoxification Model

   (b) **detoxification**: the Bart-base-detox model processes the input text, identifies toxic elements and generates a detoxified version of the text.
   (c) **output generation**: the detoxified text that maintains the original meaning as much as possible is generated.

(3) **Backtranslation**. Convert the detoxified English text back to the original minor language. Google Cloud Translation API is used to translate the detoxified English text back to the minor language.

(4) **Post-processing**. Eliminate toxic content in the translated text. The post-processing step utilizes predefined toxic keywords for each minor language. This step involves:
   (a) **keyword matching**: the detoxified minor language text is scanned for toxic keywords using the predefined dictionaries.
   (b) **keyword elimination**: toxic keywords are either removed or replaced with neutral word.
   (c) **output**: the detoxified text is generated.

3. Experimental Setup and Results

3.1. Datasets

The experimental dataset is provided by the organizers and contains a corpus of toxic sentences in nine different languages (i.e., English (en), Spanish (es), German (de), Chinese (zh), Arabic (ar), Hindi (hi), Ukrainian (uk), Russian (ru), and Amharic (am)). These datasets are divided into Development and Test phases. The former contains 400 pairs for each of the 9 languages and the latter includes 600 toxic sentences per language. Table 1 presents a summary of the datasets.

Table 1
Experimental Dataset Statistics

Dataset	en	es	de	zh	ar	hi	uk	ru	am
Development	400	400	400	400	400	400	400	400	400
Test	600	600	600	600	600	600	600	600	600

3.2. Experimental Setup

For the provided test data, having 5400 toxic sentences across 9 different languages, the column "neutral_sentence" is initially empty and the corresponding task is to fill it with the predicted detoxification text. For each detoxification text, we create a JSON object with two keys ("id" and "text"), where the former indicates the sequence number of the toxic sentence and the latter represents the detoxification sentence. The detoxification text is then filled into the "neutral_sentence" column.

Fine-tuning involves adjusting model hyperparameters to improve its performance on specific tasks. For the mT5 and Bart models, we fine-tune them with the training data provided for the multilingual text detoxification task. For the learning rate, we experiment with learning rates ranging from 1e-5 to 5e-4. The optimal learning rate for mT5 is 3e-5 and the optimal learning rate for Bart is 2e-5. For batch size, we test batch sizes of 8, 16, 32, and 64. mT5 achieves better trade-off between memory usage and training efficiency with a batch size of 32. Considering GPU memory constraints, a batch size of 16 is used for Bart.

The $\mathcal{J}$ score, referred to as the Joint metric, is used to evaluate the performance of the text detoxification model. $\mathcal{J}$ score combines three components: Style Transfer Accuracy (STA), Content Preservation (SIM), and Fluency (FL), and is calculated as the mean of the product of these three components per sample. As shown in equation (1). For the three metrics, STA measures the level of non-toxicity of the generated paraphrase, SIM evaluates the similarity of content between the original and detoxified text, and FL estimates the adequacy of the text and its similarity to human-written detoxified references.

$$J = \frac{1}{N}\sum_{i=1}^{N} \text{STA}(x_i) \cdot \text{SIM}(x_i) \cdot \text{FL}(x_i) \tag{1}$$

3.3. Baseline

The proposed model is compared three baseline methods, which are provided by the competition organizer.

Deletion. It eliminates toxic keywords based on a predefined dictionary for each language.

Backtranslation. It first utilizes the NLLB-600M model to translate the minor language text, then uses the English bart-base-detox model to detoxify the text, and finally translates it back to original language.

mT5. It is a multilingual version of T5, a text-to-text transformer model trained on over 100 languages for various downstream tasks.

3.4. Results

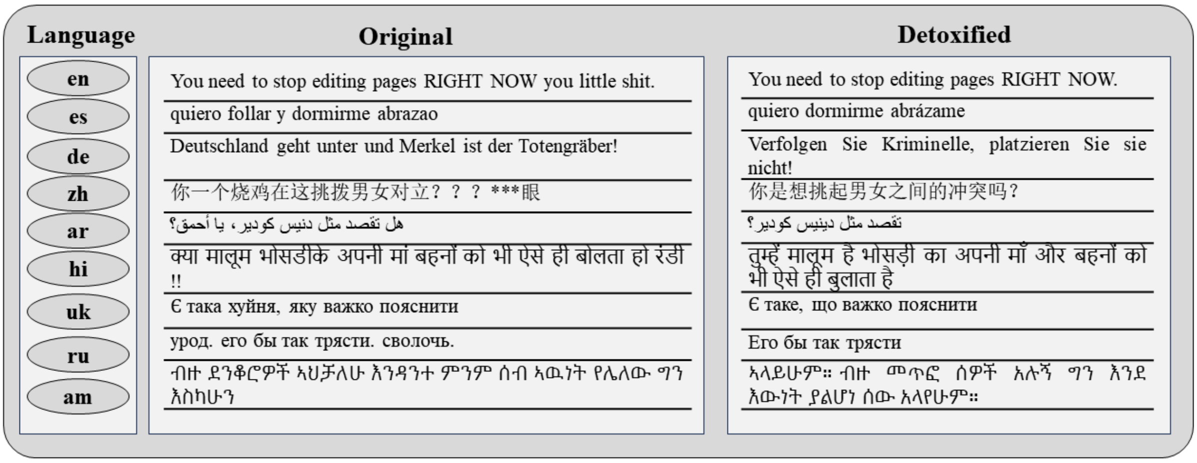
We conduct comparative experiments using the competition datasets provided by PAN@CLEF 2024. Figure 3 provides examples of detoxified sentences for the nine languages and Table 2 presents the human evaluation results of the multilingual and cross-lingual experiments in the testing phase. The first column gives different detoxification models, and the remaining nine columns represent different languages. For each language, the best result is highlighted in bold. From Table 2, we can observe that GCTP achieves the highest $\mathcal{J}$ score for Amharic text and ranked the 5th place for Chinese text. This partially demonstrates the effectiveness of GTCP. Besides, the simple *Deletion* method performs well in the Arabic, Hindi, and Ukrainian settings, which is possibly due to the lack of sufficient translation data.

The average scores in Table 2 highlight that GTCP performs well across multiple languages but faces challenges with certain languages such as Hindi. In the Hindi results, the detoxified text contains errors where offensive terms are not correctly removed or neutralized. This issue highlights the need for more effective preprocessing and post-processing steps to ensure the removal of toxic content.

Table 2
Manual Evaluation Results for 100 Randomly Selected Texts per Language

Model	en	es	de	zh	ar	hi*	uk	ru	am	average
Deletion	0.47	0.55	0.57	0.43	**0.65**	**0.65**	**0.60**	0.49	0.63	0.56
mT5	0.68	0.47	**0.64**	0.43	0.63	0.60	0.42	0.40	0.61	0.54
Backtranslation	0.73	**0.56**	0.34	0.34	0.42	0.33	0.23	0.22	0.54	0.41
GTCP (Ours)	**0.73**	0.52	0.01	**0.56**	0.49	*0.49	0.42	**0.68**	**0.72**	0.51

Note: * indicates the detoxified text for Hindi is incorrect.



Language	Original	Detoxified
en	You need to stop editing pages RIGHT NOW you little shit.	You need to stop editing pages RIGHT NOW.
es	quiero follar y dormirme abrazao	quiero dormirme abrázame
de	Deutschland geht unter und Merkel ist der Totengräber!	Verfolgen Sie Kriminelle, platzieren Sie sie nicht!
zh	你一个烧鸡在这挑拨男女对立？？？***眼	你是想挑起男女之间的冲突吗？
ar	هل تقصد مثل دنيس كودير، يا أحمق؟	تقصد مثل دينيس كودير؟
hi	क्या मालूम भोसडीके अपनी मां बहनों को भी ऐसे ही बोलता हो रंडी !!	तुम्हें मालूम है भोसड़ी का अपनी माँ और बहनों को भी ऐसे ही बुलाता है
uk	Є така хуйня, яку важко пояснити	Є таке, що важко пояснити
ru	урод. его бы так трясти. сволочь.	Его бы так трясти
am	ብዙ ደንቆሮዎች ኣህቻለሁ እንዳንተ ምንም ሰብ ኣዉነት የሌለው ግን እስካሁን	ኣላይሁም። ብዙ መጥፎ ሰዎች አሉኝ ግን እንደ እውነት ያልሆነ ሰው አላየሁም።

Figure 3: Examples of Detoxified Sentences for Different Languages

4. Conclusion

Text detoxification is a challenging task, especially for small languages where there is no sufficient data in training a powerful detoxification model. To this end, we in this study propose a detoxification model that utilizes the Google Cloud Translation and post-processing. Specifically, motivated by the transfer learning, we first use Google Cloud Translation to transfer the target domain (i.e., minor language text) into source domain (i.e., English language text) and then perform detoxification. Afterwards, the text is translated back into minor language text and post-processed using predefined toxic keywords. Finally, the proposed model is compared with three baseline methods on the competition datasets. Results show its effectiveness.

References

[1] K. Wang, J. Yang, H. Wu, A survey of toxic comment classification methods, arXiv preprint arXiv:2112.06412 (2021).

[2] 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.

[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] 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, in: Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval), 2022, pp. 90–101.

[5] 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.

[6] 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.

[7] 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).

[8] 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).

[9] S. Pletenev, Between denoising and translation: Experiments in text detoxification (2022).

[10] E. Totmina, Detoxification of russian texts based on combination of controlled generation using pre-trained rugpt3 and the delete method, in: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue, volume 21, 2022, pp. 1167–1174.

[11] 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.

[12] 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 (2022).

[13] 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).