Team Detox at TextDetox CLEF 2025: Multilingual Text Detoxification using LLM

Notebook for the PAN Lab at CLEF 2025

Gopala Krishna Nuthakki[1,*], Lekkala Sai Teja[2] and Atul Mishra[1]

[1]Computer Science & Engineering, BML Munjal University, Haryana, India
[2]Computer Science & Engineering, National Institute of Technology, Silchar, India

Abstract

Toxic online language poses a severe threat to the safety and inclusivity of users on many online websites. This work is a part of PAN at CLEF 2025 shared task named Multilingual Text Detoxification (TextDetox) shared task 2025, which tries to convert toxic texts into non-toxic ones while preserving semantic meaning in different languages that range from being high-resource to underrepresented ones. The dataset used in this task consists of toxic sentences in 15 languages from around the globe such as English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian, Amharic, Italian, French, Hebrew, Hinglish, Japanese and Tatar, as provided by the CLEF PAN-25 initiative. Our method, applicable to all 15 languages, the approach begins by identifying and masking toxic words in input sentences. The original and masked versions are then provided to large language models (LLMs) to generate detoxified outputs that retain the intended meaning while eliminating offensive language. These experiments are evaluated based on style accuracy, semantic preservation, and fluency. The study show competitive results across multiple languages, highlighting the effectiveness of a hybrid approach in multilingual style transfer tasks. Our results further go toward inclusive and robust moderation tools that allow safer communication in multilingual digital spaces. All our codes can be seen on GitHub[1]."

Keywords

PAN 2025, Large Language Models, Text Detoxification, Multilingual

1. Introduction

Online interactions are becoming multilingual, and while this global reach is of immense value, it also poses the challenge of moderating offensive and toxic content [1] in linguistic and cultural contexts. Maintaining online discussions respectfully and safely requires methods that can effectively detoxify language without watering down the intended meaning of a message.

The Multilingual Text Detoxification (TextDetox) 2025 [2] shared task entails translating toxic user-generated content into non-toxic content without sacrificing the semantic meaning of the source content in languages. This task is hosted as part of a broader effort to combat online toxicity, moving beyond conventional content moderation strategies that rely on blocking or removing harmful texts. Instead, the goal is to proactively rewrite toxic content, preserving its core message while eliminating offensive or obscene language. The task focuses on explicit toxicity, which includes direct use of obscene or rude lexicons where neutral content can still be extracted and preserved.

One of the main challenges in this area is the variability in toxic phrasing across languages and dialects. The vast majority of languages lack labeled data for this issue, and supervised learning becomes unfeasible. In addition, translating detoxed content often loses its meaning or gets misinterpreted as culturally unpleasant.

To address these challenges, The proposed study used a hybrid approach that combines rule-based masking with model-guided generation. First, toxic spans were identified using keyword-based lexicons. Such lexicons were subsequently masked to reduce generation bias. By feeding the original and

[1]https://github.com/gopalkrishna2004/CLEF-PAN-2025-Multilingual-Text-Detoxification

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.
sivagopalkrishna04@gmail.com (G. K. Nuthakki); lekkalad_ug_22@cse.nits.ac.in (L. S. Teja); atul.mishra@bmu.edu.in (A. Mishra)

masked sentences into multilingual large language models, we ensured that the detoxified outputs were semantically aligned with the input and effectively removed offensive content. This two-input method allowed the model to better identify contextual nuances, especially in code-mixed and low-resource languages, and improved the outputs.

2. Dataset Description

The dataset, acquired via CLEF 2025 PAN [3] , contains 400 parallel sentence pairs containing a toxic and detoxified version for 9 languages: English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian, and Amharic. In addition, the dataset also includes a Toxic Keywords List for 15 languages mentioned above in the shared task. This resource captures commonly occurring offensive or toxic terms and phrases, serving both as lexical guidance for detoxification models and as a benchmark for evaluating toxic span identification. Furthermore, toxic span annotations are available for the same 9 languages with parallel data, enabling more fine-grained analysis.

3. Methodology

The section outlines the specific design and development approach used for developing our multilingual text detoxification system. We explain the architecture, low-level techniques behind identifying and cleaning up toxic text to create detoxified versions of them, and how implementation strategies hold across languages. The objective of the study is to develop a generalizable system in support of several languages and cleaning up toxic input without sacrificing any original meaning or context.

3.1. Overall System Architecture

The system is designed to accommodate a multilingual detoxification pipeline in which input text in toxic form is converted into its non-toxic equivalent for different languages. The system starts with a toxic sentence submitted by the user in one of the languages supported. This input is fed to a modular detoxification engine that uses several detoxification models concurrently. Each model processes the input independently and produces its detoxed equivalent. These outputs are then gathered and assessed to compare fluency, factuality consistency, and reduction of toxicity. The architecture was made such a way that it could be easily extended with more LLMs.

The framework has multiple functional layers. The Input Layer takes raw toxic sentences without preprocessing. In the Masking Layer, toxic words of the sentence are detected automatically and masked to maintain sentence structure with a focus on objectionable content. The Model Layer also takes the original and masked sentences as inputs to multilingual LLMs. The models give detoxified versions of text that try to remove toxicity while maintaining the original semantic intent.

3.2. Detoxification Models and Techniques

The toxic dataset comprises 9,000 instances in the test set. For every toxic sentence, first of all, the toxic words are deleted, and they get replaced with the token ([MASK]). This gives a masked version of the original sentence, keeping the structure intact without offensive content.

Both the original toxic sentence and its censored counterpart are fed into large language models (LLMs) with few-shot prompting. The study used multilingual instruction-following LLMs such as GPT-4o-mini[4], which is prompted to generate detoxified outputs that retain the semantic integrity of the original sentence while eliminating toxicity. These models prompted in a way to produce grammatically coherent, contextually correct, and non-toxic completions.

This hybrid approach produces effective detoxification across both high-resource and low-resource languages, without depending upon supervised fine-tuning or parallel corpora. Instead, it uses the generative capabilities of LLMs through few-shot prompting and contextual understanding. Refer to Fig.1 and Fig.2 for an overview of the masking and generation pipeline.

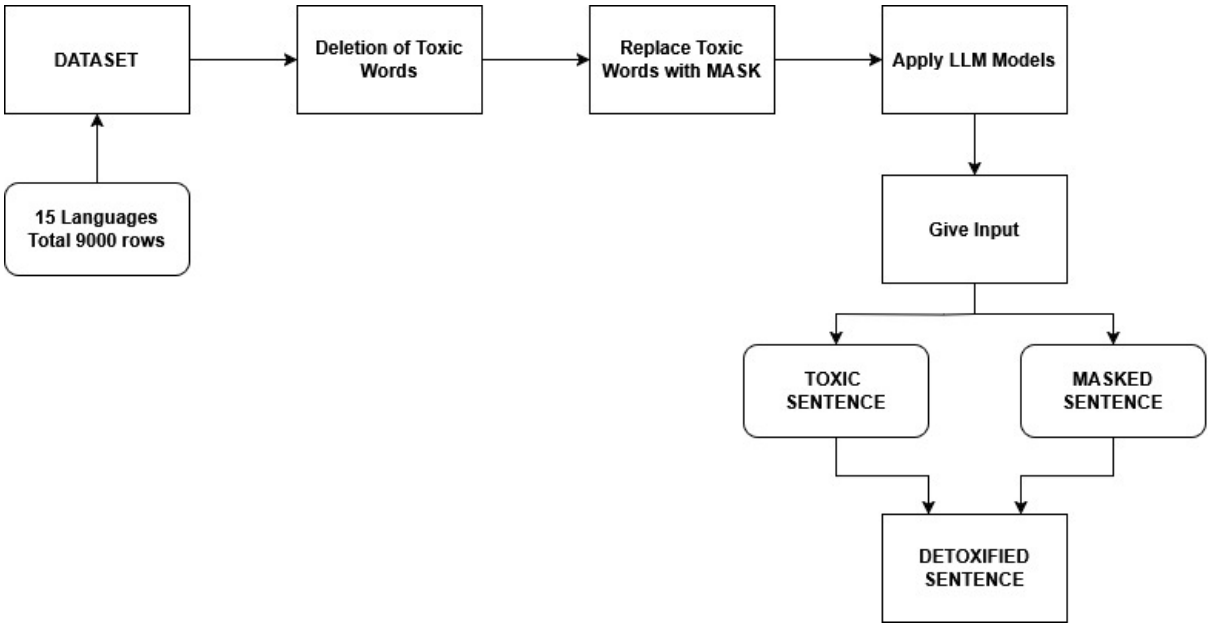

Figure 1: System Diagram of the Process

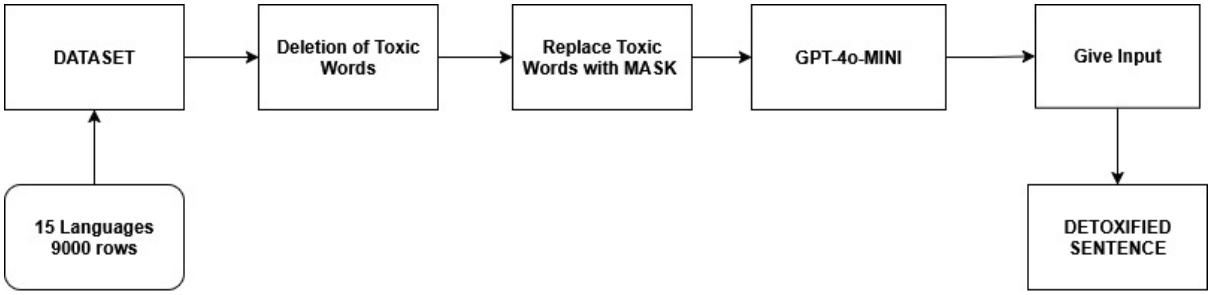

Figure 2: Model Architecture of Supervised and Unsupervised data

3.3. Implementation

There are two major steps of our implementation as follows:

- **Toxic Word Deletion(lexicon-based):** This defines a class called Detoxification, which uses a multilingual word list to filter out toxic words. The word list can either be loaded from a local file or taken from the multilingual-toxic-lexicon dataset [5]. For non-Chinese text, the sentence is split into words using spaces. Toxic words are removed, and the remaining words are joined back together. For Chinese text, the sentence is first broken into words using jieba, then toxic words are removed. In both cases, the harmful words are replaced with a general token [MASK], and the cleaned sentences are saved for the next step.
- **Few-shot LLM-based Style Transfer Stage (Masked Completion):** The original toxic sentences, along with their corresponding masked versions, are provided to instruction-following large language models (LLMs) such as GPT-4o-mini, using few-shot prompting. The system prompt includes clear formatting instructions and example cases to guide the model in replacing the [MASK] token with appropriate detoxified content. The prompt consists:
  - Toxic sentence
  - Masked sentence

– Few shot prompting to generate the detoxified output in the original language

The model takes these input texts and generates sentences that are grammatically correct, contextually relevant, and not offensive.

4. Evaluation Metrics

To evaluate the quality of detoxified outputs across languages, TextDetox 2025 shared task used three major evaluation metrics : Style Transfer Accuracy, Content Preservation, and Fluency.

4.1. Automatic Evaluation

4.1.1. Style Transfer Accuracy

Style Transfer Accuracy assesses whether the output successfully transfers a toxic sentence into a non-toxic sentence. For this, we employed a binary toxicity classifier on the basis of the XLM-Roberta-Large model [6] that was specifically fine-tuned for toxicity detection. This metric measures how well the detoxification model modifies the style of the input while removing toxic language.

4.1.2. Content Preservation

Content Preservation is the measure of how closely the semantic meaning of the original toxic sentence is to the detoxified sentence. This is calculated as cosine similarity between LaBSE [7] embeddings of input and output sentences. The similarity score is higher, the more the detoxified sentence maintains the original meaning.

4.1.3. Fluency

Fluency is utilized to approximate the degree to which the produced sentences are natural, grammatically sound, and coherent. For this purpose, the xCOMET[8] model is utilized, which has been shown to have a high correlation with human fluency judgment of detoxified text. The COMET machine translation models are used as a robust proxy to evaluate the adequacy and linguistic quality of the output.

$$\begin{aligned} J = & X_{comet\_fluency}(input, output_{gold}, output_{generated}) \\ & \times (0.4 \times Similarity(input, output_{generated}) \\ & \quad + 0.6 \times Similarity(output_{gold}, output_{generated})) \\ & \times STA \end{aligned} \tag{1}$$

$$STA = \frac{sta\_scores + \frac{\sum compared\_scores}{len(compared\_scores)}}{2} \tag{2}$$

Where:

$$sta\_scores = classifier\_prob\_neutral(output_{generated}) \tag{3}$$

$$compared\_scores = sta\_scores \leq ref\_sta\_scores \tag{4}$$

$$ref\_sta\_scores = classifier\_prob\_neutral(output_{gold}) \tag{5}$$

4.2. LLM-as-a-Judge Evaluation

For this shared task, the evaluation is also carried out in another way, namely LLM-as-a-Judge, a popular tool in recent times for the evaluation. In our submissions, this framework is used for the evaluation, using a *LLaMA 3.1-8B-Instruct* fine-tuned version, which is trained on human-annotated pairwise comparisons that have been taken from the TextDetox 2024 dataset. This fine-tuned model

evaluates the results by comparing them in pairs, which would be more human-aligned judgments than traditional automatic metrics. For the fluency evaluation, the xCOMET model is used as always due to its strong correlation with human fluency assessments.

5. Results

The sections demonstrate a detailed assessment of the multilingual text detoxification models on their performance in the 15 languages of the Multilingual ParaDetox dataset. The task offer both quantitative outcomes based on traditional evaluation measures as well as qualitative examples of model behavior across varied linguistic and stylistic scenarios. The discussion further comments on model generalizability implications, and model effectiveness across low-resource languages.

5.1. Quantitative Results

Table 1 and Table 2 show the automatic evaluation 4.1 results, while Table 3 and Table 4 shows the LLM-as-a-Judge Evaluation 4.2 results of the hybrid LLM approach using GPT 4o-mini and the baseline methods. The detoxification itself involved feeding GPT 4o-mini a toxic sentence and its masked version (where toxic terms are replaced by placeholders), then asking it to produce a non-toxic one. The outputs were evaluated based on the same three measures: Style Transfer Accuracy (STA), Content Preservation (SIM), and Fluency (FL), with the same procedures outlined above. The final score for every language and model was obtained by applying the multiplicative formula STA × SIM × FL, and the mean score across supervised and unsupervised languages was reported to indicate the effectiveness of the model in unsupervised detoxification tasks.

Table 1
Automatic Evaluation results on Unsupervised set

Model	Language(s)	Average Score
hybrid gpt-4o-mini	**6 Unsupervised languages**	**0.595**
baseline_gpt4	6 Unsupervised languages	0.595
baseline_mt0	6 Unsupervised languages	0.572
baseline_gpt4o	6 Unsupervised languages	0.535
baseline_delete	6 Unsupervised languages	0.510
baseline_o3mini	6 Unsupervised languages	0.484
baseline_duplicate	6 Unsupervised languages	0.482
baseline_backtranslation	6 Unsupervised languages	0.342

Table 2
Automatic Evaluation results on the supervised set

Model	Language(s)	Average Score
baseline_mt0	9 supervised languages	0.675
baseline_gpt4	9 supervised languages	0.637
hybrid gpt-4o-mini	**9 supervised languages**	**0.611**
baseline_o3mini	9 supervised languages	0.562
baseline_gpt4o	9 supervised languages	0.560
baseline_delete	9 supervised languages	0.536
baseline_backtranslation	9 supervised languages	0.481
baseline_duplicate	9 supervised languages	0.475

Table 3
LLM-as-a-Judge Evaluation results across 6 Unsupervised languages. **Team Detox**'s results are highlighted.

Rank	Team	Average	it	ja	he	fr	tt	hin
1	**golden annotation**	**0.828**	**0.893**	**0.904**	**0.783**	**0.724**	**0.780**	**0.887**
2	Team ReText.AI Team	0.722	0.823	0.805	0.657	0.860	0.583	0.606
3	ducanhhbtt	0.720	0.842	0.820	0.681	0.889	0.495	0.592
4	**Team Detox**	**0.704**	**0.812**	**0.784**	**0.631**	**0.843**	**0.575**	**0.578**
11	baseline_gpt4	0.662	0.790	0.779	0.578	0.865	0.438	0.524
14	baseline_mt0	0.641	0.749	0.711	0.501	0.793	0.598	0.494
24	baseline_o3mini	0.559	0.748	0.661	0.497	0.826	0.209	0.411
27	baseline_gpt4o	0.526	0.697	0.680	0.370	0.718	0.327	0.363
28	baseline_delete	0.525	0.628	0.443	0.496	0.576	0.521	0.486
30	baseline_duplicate	0.429	0.455	0.442	0.407	0.460	0.421	0.387
31	baseline_backtranslation	0.254	0.333	0.147	0.349	0.503	0.054	0.139

Table 4
LLM-as-a-Judge evaluation results across 9 supervised languages. **Team Detox**'s results are highlighted.

Rank	Team	Average	en	es	de	zh	ar	hi	uk	ru	am
1	**golden annotation**	**0.820**	**0.846**	**0.783**	**0.930**	**0.716**	**0.838**	**0.888**	**0.807**	**0.828**	**0.742**
2	Team MetaDetox	0.812	0.893	0.823	0.919	0.813	0.826	0.785	0.791	0.829	0.626
3	ducanhhbtt	0.798	0.871	0.797	0.919	0.796	0.814	0.762	0.785	0.827	0.614
4	Team ReText.AI Team	0.775	0.794	0.765	0.888	0.783	0.790	0.773	0.791	0.792	0.597
16	**Team Detox**	**0.722**	**0.691**	**0.757**	**0.819**	**0.699**	**0.718**	**0.701**	**0.742**	**0.792**	**0.580**
17	baseline_gpt4	0.715	0.858	0.800	0.807	0.654	0.686	0.647	0.723	0.778	0.482
24	baseline_o3mini	0.676	0.893	0.796	0.747	0.652	0.595	0.609	0.663	0.711	0.421
27	baseline_gpt4o	0.580	0.770	0.694	0.533	0.482	0.539	0.490	0.608	0.615	0.486
29	baseline_delete	0.558	0.453	0.543	0.564	0.630	0.610	0.566	0.577	0.583	0.499
30	baseline_backtranslation	0.458	0.743	0.466	0.479	0.231	0.442	0.395	0.256	0.689	0.425
31	baseline_duplicate	0.432	0.370	0.451	0.479	0.429	0.446	0.432	0.455	0.450	0.380
32	Dorevain	0.346	0.838	0.407	0.417	0.170	0.337	0.288	0.152	0.201	0.305

5.2. Qualitative Results

Cross-lingual examples are presented in Table 5 to demonstrate the project's capability to rephrase toxic inputs while preserving their original meaning and improving overall fluency. For highly toxic inputs, some outputs became vague or overly neutral. Using masked inputs helped the model focus better on toxic segments

Table 5
Examples of Toxic, Masked, and Detoxified Text

Toxic	Masked	Detoxified
do people actually believe this shit	do people actually believe this [MASK]	do people actually believe this nonsense
don't be such a nitpicky cunt, it's bad for	don't be such a nitpicky [MASK], it's bad for	don't be such a nitpicky person, it's bad for

6. Conclusion

This paper provides a study of multilingual text detoxification, introduced as part of the PAN 2025 shared task. Our hybrid method that integrates lexicon based toxic word masking with few-shot prompting of multilingual LLMs such as GPT-4o-mini showed robust detoxification performance on both high-resource domain as well as low-resource languages compared to baseline methods. Semantic

consistency was preserved during the removal of explicit toxic content, as confirmed by strong results in style transfer accuracy, semantic preservation, and fluency. These outcomes were validated using both automated evaluation methods and assessments by large language models acting as judges.

The findings indicate that hybrid models using large generative models are able to generalize across widely varying linguistic patterns and cultural contexts even without large parallel corpora. This research provides the foundation for developing robust, scalable, and inclusive content moderation tools that can engage multilingual digital societies. Future work will focus on enhancing domain adaptation, managing implicit toxicity, and scaling the model to real-world applications.

All models were evaluated using three carefully selected metrics to comprehensively assess the effectiveness of detoxification: Style Transfer Accuracy (quantified with XLM-RoBERTa fine-tuned for toxicity detection), Content Preservation (cosine similarity of LaBSE embeddings), and Fluency (with the xCOMET model that agrees well with hu- man ratings on text quality) and also through LLM-as-a-Judge. A combined score function with these metrics enabled us to compare models fairly across languages.

Declaration on Generative AI

The author(s) used Grammarly, ChatGPT, and Gemini for language editing and LaTeX table formatting. The content was reviewed and finalized by the author(s), who take full responsibility for it.

References

[1] P. Madhyastha, A. Founta, L. Specia, A study towards contextual understanding of toxicity in online conversations, Natural Language Engineering 29 (2023) 1538–1560. doi:10.1017/S1351324923000414.

[2] D. Dementieva, V. Protasov, N. Babakov, N. Rizwan, I. Alimova, C. Brune, V. Konovalov, A. Muti, C. Liebeskind, M. Litvak, D. Nozza, S. Shah Khan, S. Takeshita, N. Vanetik, A. A. Ayele, F. Schneider, X. Wang, S. M. Yimam, A. Elnagar, A. Mukherjee, A. Panchenko, Overview of the multilingual text detoxification task at pan 2025, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025.

[3] J. Bevendorff, D. Dementieva, M. Fröbe, B. Gipp, A. Greiner-Petter, J. Karlgren, M. Mayerl, P. Nakov, A. Panchenko, M. Potthast, A. Shelmanov, E. Stamatatos, B. Stein, Y. Wang, M. Wiegmann, E. Zangerle, Overview of PAN 2025: Voight-Kampff Generative AI Detection, Multilingual Text Detoxification, Multi-Author Writing Style Analysis, and Generative Plagiarism Detection, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.

[4] OpenAI, Gpt-4o mini: Advancing cost-efficient intelligence, 2024. URL: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, accessed: 2025-05-31.

[5] TextDetox, Multilingual toxic lexicon, https://huggingface.co/datasets/textdetox/multilingual_toxic_lexicon, 2023. Accessed: 2025-05-31.

[6] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, 2020. URL: https://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[7] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic bert sentence embedding, 2022. URL: https://arxiv.org/abs/2007.01852. arXiv:2007.01852.

[8] N. M. Guerreiro, R. Rei, D. van Stigt, L. Coheur, P. Colombo, A. F. T. Martins, xcomet: Transparent machine translation evaluation through fine-grained error detection, 2023. URL: https://arxiv.org/abs/2310.10482. arXiv:2310.10482.