HF_Detox at TextDetox CLEF 2025: Prompt-Driven Multilingual Detoxification

Notebook for PAN at CLEF 2025

Humaira Farid[1,†], Zainab Ahmad[2,†], Ahmad Mahmood[2,†] and Iqra Ameer[3,*,†]

[1]Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802, USA
[2]Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación(CIC), Mexico City, Mexico
[3]Division of Science and Engineering, The Pennsylvania State University, Abington, PA, 19001, USA

Abstract

The text detoxification task aims to automatically transform toxic or offensive sentences into neutral, semantically equivalent paraphrases. In this study, we evaluated a lightweight, prompt-based chain-of-thought approach using OpenAI's GPT-4o-mini on the PAN 2025 multilingual ParaDetox benchmark, encompassing 15 typologically diverse languages. Without any fine-tuning, our method leverages a single system instruction with a few-shot setup. During the initial evaluation, our approach achieved top-3 Joint (J) scores in six languages, most notably French (J = 0.775, rank 2) and Hebrew (J = 0.613, rank 1), and ranked in the top 10 for 9 languages overall. In the post-evaluation phase, our system ranked fourth overall in the parallel and third non-parallel data tracks, achieving average J scores of 0.768 and 0.718, respectively. Strong results were observed in English (J = 0.775), Spanish (J = 0.814), Japanese (J = 0.819), and Hebrew (J = 0.671), achieving the second position, demonstrating our method's robustness in both high-resource and morphologically diverse languages. These findings underscore the potential of multilingual large language models, guided by carefully designed prompts, to serve as plug-and-play detoxifiers even in the absence of task-specific fine-tuning.

Keywords

PAN 2025, Multilingual Text Detoxification (TextDetox) 2025, Style Transfer, Multilingual NLP, GPT-4o-mini

1. Introduction

The rapid and widespread growth of social media platforms, including YouTube, Facebook, Twitter, and Instagram, in recent years has transformed the culture of communication and interaction around the world. These platforms generate a huge amount of user-generated content on a daily basis, providing rich data for natural language processing (NLP) research and applications [1]. However, they're also hosts of toxic and harmful speech that can lead to harassment, exclusion, and even radicalization [2].

To mitigate this, most online platforms rely on content moderation techniques such as blocking or deleting harmful posts. Although these are good countermeasures to some extent, these measures are reactive and often result in the loss of potentially meaningful content. A more proactive approach involves text detoxification, which consists of rewriting toxic content in a non-offensive manner while retaining the original meaning of the content [3].

In this paper, we tackle the Multilingual Text Detoxification task, which was proposed as part of the PAN at CLEF 2025 shared task [4]. The aim is to transform toxic sentences into neutral and non-offensive versions in a diverse range of languages (a total of 15), including high-resource languages (e.g., English, Spanish) as well as low-resource (e.g., Amharic, Hinglish, and Tata) or code-switched languages (e.g.,

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
†These authors contributed equally.
humaira.farid@gmail.com (H. Farid); zainabshaukat09@gmail.com (Z. Ahmad); ahmadmahmood447@gmail.com (A. Mahmood); iqa5148@psu.edu (I. Ameer)
https://www.abington.psu.edu/person/iqra-ameer-phd (I. Ameer)
0009-0006-7353-0362 (H. Farid); 0009-0006-4596-8973 (Z. Ahmad); 0009-0002-0755-3145 (A. Mahmood); 0000-0002-1134-9713 (I. Ameer)

Hinglish). The emphasis is on explicit toxicity, such as being rude or using swear words, rather than more implicit or subtle forms (e.g., sarcasm).

Several challenges make this task both complex and impactful. One of the main obstacles is deciding how to identify toxicity in a consistent manner across multiple languages and in diverse cultural contexts. Although some types of toxicity can be determined from explicit abusive language using lexicon-based methods, the context and construction of language, particularly across diverse linguistic families, can pose significant challenges. Another serious issue is ensuring that the original meaning of the content is retained in detoxification. Rephrasing sentences to eliminate toxic content without altering the original meaning requires deep language understanding. Additionally, the discrepancy in resource availability across different languages makes this task further challenging. While high-resource languages (e.g., English, Spanish, etc.) benefit from large-scale corpora and pre-trained models, low-resource or code-switched languages (e.g., Amharic, Hinglish) lack such support, making effective detoxification more difficult. Moreover, ensuring that the model's output is not only detoxified but also fluent, natural-sounding, and culturally appropriate adds further complexity [5].

In this study, we propose a truly multilingual, prompt-driven approach to text detoxification using only the GPT-4o-mini model. Rather than fine-tuning separate encoder–decoder or decoder-only architectures, we leverage in-context learning: each toxic input is prefaced with a single fixed instruction. This simple setup scales seamlessly to all 15 languages in the PAN 2025 multilingual ParaDetox benchmark. Despite its lightweight nature, without any additional fine-tuning or model assembly, our method achieves competitive performance. The results demonstrate that a well-crafted chain-of-thought prompt alone can rival more complex, fully fine-tuned systems across diverse linguistic settings.

2. Related Work

Text detoxification has attracted significant attention from researchers working across multiple languages and styles. Early efforts focused on lexicon and rule-based filtering [6] to identify and replace or remove abusive terms, but these methods proved less effective in context-sensitive scenarios. This led to data-driven paraphrasing approaches [7], most notably back-translation, that could rephrase toxic inputs more accurately. Later, neural style-transfer models [8] were introduced to disentangle content and toxicity signals to generate cleaner text. State-of-the-art methods fine-tune pretrained large language models [9], which achieved superior fluency, semantic fidelity, and toxicity reduction.

In 2024, PAN also introduced a shared task on multilingual text detoxification [10]. The task was to challenge participants to convert toxic or abusive text into non-toxic paraphrases. The dataset comprised text instances in nine different languages, including English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian, and Amharic. In the development phase, only English and Russian parallel data were available, while for testing, systems were evaluated on full parallel corpora consisting of nine languages. Submissions were assessed both automatically and through human crowd-sourcing. The evaluation was performed using style transfer accuracy, semantic preservation (LaBSE cosine similarity), and fluency (ChrF1 against human references). Baselines included simple duplications, delete keyword heuristics, back translation, and fine-tuned mT5. Analysis showed that multilingual LLMs (few/zero-shot and fine-tuned variants mT0-XL and LLaMa3) generally outperformed unsupervised methods, though performance varied significantly by language.

In human evaluation, Team SomethingAwful [11] claimed the first place by using an "uncensored" LLaMa3-70B[1] model with a few-shot prompting (10 exemplars per language) and a fine-tuned mT0-XL for Amharic, it achieved the highest average Joint (J) score of 0.774, with its best performance in German (J = 0.889) and Spanish (J = 0.834). Team SmurfCat [12] followed in second place by fine-tuning mT0-XL [13] on the shared task data and applying Odds-Ratio Preference Optimization (ORPO) [14] at inference, obtained an average J = 0.741 and peaking on Ukrainian (J = 0.840) and Arabic (J = 0.819). In third place, Team VitalyProtasov [15] trained mT0-large [13] with language-specific data filtering before fine-tuning, reaching an average J = 0.723 and demonstrating particular strength in Hindi (J = 0.788).

[1] https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md Last visited: 27/05/20225

On the automatic leaderboard, Team SmurfCat [12] again led with an average J = 0.523 by leveraging mT0-XL plus ORPO. Team lmeribal secured second place with an average J = 0.515; its top result was in Ukrainian (J = 0.686), suggesting a robust multilingual fine-tuning strategy despite limited method details. In third place, Team nikita.sushko [16] fine-tuned mT0-XL in two stages (parallel then synthetic data filtered by LaBSE similarity and toxicity), with an additional "delete" post-processing step, achieving an average J = 0.465 and the highest scores in Ukrainian (J = 0.668) and English (J = 0.553).

As the literature depicts, PAN 2024's TextDetox task highlighted the effectiveness of multilingual LLMs, especially when combined with prompt engineering or lightweight preference-optimization techniques, in outperforming unsupervised baselines. While top systems achieved impressive fluency and content preservation, their performance still varied by language and toxicity category, underscoring ongoing challenges in contextual nuance and low-resource adaptation.

3. Data

We use the ParaDetox parallel corpus from the Multilingual Text Detoxification task at PAN 2025, available on Hugging Face [17, 18].[2] The dataset contains sentence-level toxic–detoxified pairs in nine languages: English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian, and Amharic, which serves as a training set. For all nine languages, there is a pair of 400 toxic-neutral sentences. Similarly, the test set consists of 15 languages, including English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian, Amharic, Italian, French, Hebrew, Hinglish, Japanese, and Tatar. Example for English, Spanish and Genrman are presentd in 1. For all 15 languages, there are 600 toxic sentences. The table 2 provides a detailed overview of the dataset.

Table 1
Examples of Toxic and Detoxified Sentences Across Different Languages

Language	Toxic Sentence	Detoxified Sentence
English	You're so stupid, I can't believe anyone listens to you.	I think your idea might need more clarity.
Spanish	Eres tan estúpido que no entiendo cómo alguien te escucha.	Creo que tu idea necesita un poco más de claridad.
German	Du bist so dumm, ich kann nicht glauben, dass dir jemand zuhört.	Ich denke, deine Idee könnte klarer formuliert werden.

Table 2
Dataset Statistics

Category	Training Set	Testing Set
Number of Languages	9	15
Instances per Language	400	600
Total Instances	3,600	9,000

4. Methodology

Our approach relies exclusively on prompt engineering with the GPT-4o-mini model, avoiding any full fine-tuning due to policy restrictions on toxic content. We explore a range of prompting techniques, from simple zero-shot instructions to detailed chain-of-thought prompts, and compare against a back-translation baseline. Figure 1 presents the overview of the methodology that how the toxic sentence is

[2] https://huggingface.co/datasets/textdetox/multilingual_paradetox

getting transformed into a detoxified one without changing its entire meaning, but by altering the toxic words or phrases using our proposed methodology.

4.1. Zero-Shot Prompting

In the zero-shot setting, we present only a high-level instruction to the model:

> [You are a wonderful text detoxification assistant, for each given query substitute the toxic word with non-toxic word.]

Despite its minimal guidance, this setup leverages GPT-4o-mini's extensive multilingual pre-training and establishes a strong baseline for detoxification.

4.2. Chain-of-Thought Prompting

To improve the handling of subtle or context-dependent toxicity, we design a chain-of-thought prompt that walks the model through the detoxification process in steps: Initially, identifying the toxic term, then proposing a neutral substitute, and finally verifying that the meaning is preserved:

> [You are a multilingual detoxification assistant. When the user sends you a sentence, follow this process: 1. "Examine the sentence for toxicity." 2. "Identify which word(s) are toxic or dirty." 3. "For each toxic word, think of a neutral synonym that preserves the original meaning." 4. "Replace the toxic word(s) with the chosen synonym(s)." 5. "Review the new sentence to ensure it's fluent and nothing else has changed." 6. "Now provide the final non-toxic paraphrase as a plain string." Do not add or remove anything besides toxic words. If the sentence is already non-toxic, simply repeat it unchanged.]

This explicit reasoning path helps the model produce more accurate and nuanced paraphrases.

4.3. Fine-Tuning Considerations

We evaluated OpenAI's fine-tuning API on our parallel toxic–neutral dataset, but we were unable to proceed due to policy constraints around harmful content. This reinforces the value of in-context learning: with carefully crafted prompts alone, our method achieved first place in Hebrew and strong performance in French without updating any model parameters.

4.4. Back-Translation Baseline

As a comparative baseline, we implemented a back-translation pipeline: translating the input into English, detoxifying in English, then translating back to the original language. Although this approach had a sound concept, it proved computationally expensive and underperformed compared to our prompt-driven techniques.

Overall, our experiments demonstrate that strategic prompt design with GPT-4o-mini offers a lightweight yet powerful solution for multilingual text detoxification across 15 languages.

5. Results and Analysis

This section presents the performance of our proposed technique on the PAN 2025 ParaDetox test set, as well as the outcomes from the subsequent post-evaluation phase. We report and analyze the Joint (J) scores and leaderboard rankings of GPT-4o-mini across all 15 target languages, highlighting trends, strengths, and limitations observed during both evaluation stages.

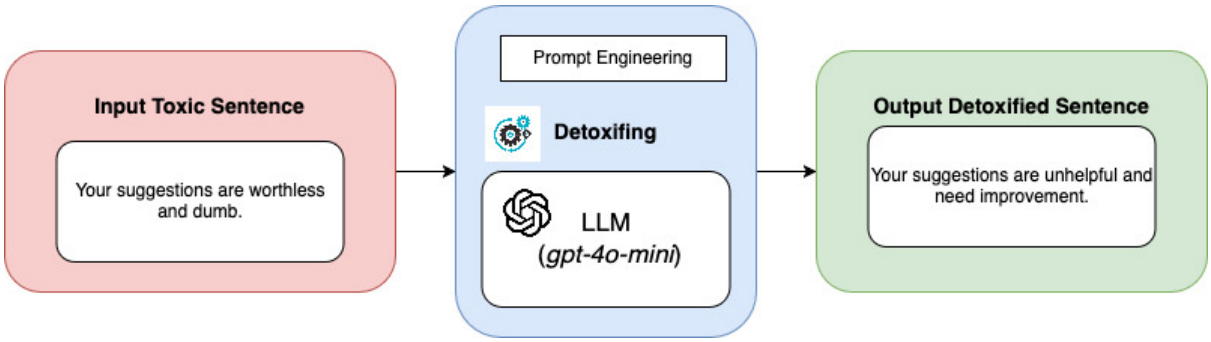

Figure 1: Overview of Methodology

5.1. Initial Results

Table 3 reports the Joint (J) scores and leaderboard ranks for GPT-4o-mini across all 15 languages in the PAN 2025 ParaDetox initial test set. The results show that GPT-4o-mini achieves its strongest performance in *French* (J = 0.775, rank 2) and *Hebrew* (J = 0.613, rank 1), indicating particularly effective in-context detoxification in these languages. High-resource European languages such as *English* (J = 0.727, rank 3) and *Spanish* (J = 0.705, rank 3) also perform well, suggesting the model's pretraining is most robust for widely represented languages.

Mid-tier scores appear on *Japanese* (J = 0.653, rank 4) and *Hindi* (J = 0.600, rank 6), indicating moderate success in East Asian languages. Notably, *German* (J = 0.699, rank 8) and *Russian* (J = 0.708, rank 7) achieve competitive scores but slightly lower ranks, perhaps due to stronger baselines from specialized systems.

Performance declines in lower-resource or morphologically complex languages: *Arabic* (J = 0.585, rank 11),*Italian* (J = 0.673, rank 12) and *Amharic* (J = 0.394, rank 12) rank lower, reflecting challenges in script variation and fewer pretraining exemplars. The poorest result is on *Hinglish* (J = 0.296, rank 17), underscoring the difficulty of mixed-script, code-switched text. These patterns suggest that while few-shot in-context learning with GPT-4o-mini excels for high- and mid-resource languages, further adaptation or fine-tuning may be required to close gaps in low-resource settings.

Table 3
Test Set Joint scores (J) and leaderboard ranks for GPT-4o-mini detoxification across 15 languages.

Language	J Score	Rank
English	0.727	3
Spanish	0.705	3
German	0.699	8
Chinese	0.506	11
Arabic	0.585	11
Hindi	0.600	6
Ukrainian	0.729	8
Russian	0.708	7
Amharic	0.394	12
Italian	0.673	12
Japanese	0.653	4
Hebrew	0.613	1
French	0.775	2
Tatar	0.439	13
Hinglish	0.296	17

5.2. Post Evaluation Results

This section presents the results for the post-evaluation phase for the parallel and non-parallel dataset.

5.3. Post Evaluation Results: Parallel Data Track

The post-evaluation Joint (J) scores and rankings for 9 languages with parallel data are presented in Table 4. Our team ranked fourth overall, along with other teams, achieving an average score of **0.768**. Notably, we obtained competitive scores in four key languages: *English* (0.888), ranked second; *Spanish* (0.814), ranked second; *German* (0.912), ranked third; and *Hindi* (0.731) ranked third. Despite lower scores in traditionally challenging languages like *Amharic* (0.526), our system consistently performed well in most of the languages. The results highlight our model's robustness in handling both morphologically diverse and high-resource languages.

Table 4
Post-evaluation Joint scores (J) and ranks for 9 languages with parallel data.

Language	J Score	Rank
Amharic	0.526	12
Arabic	0.788	9
German	0.912	3
English	0.888	2
Spanish	0.814	2
Hindi	0.731	3
Russian	0.805	9
Ukrainian	0.747	13
Chinese	0.724	7
Average	**0.768**	

5.4. Post Evaluation Results: Without Parallel Data Track

The post-evaluation Joint (J) scores and ranks for the 6 languages without parallel data, are presented in Table 5. Our team achieved an average score of **0.718** and secured the third position, which shows the robustenss of our model for unseen languages. Our method demonstrated strong performance in high-resource languages like Japanese (0.819) and Hebrew (0.883), placing second rank.

Although performance in languages like Tatar (0.511) and Hinglish (0.586) was relatively lower, these results reflect the overall difficulty associated with such low-resource or morphologically complex languages. However, consistent mid- to top-tier performance in most categories validates the effectiveness of our detoxification strategy.

Table 5
Post-evaluation Joint scores (J) for 6 languages without parallel data.

Language	J Score	Rank
Italian	0.819	5
Japanese	0.819	2
Hebrew	0.671	2
French	0.883	5
Tatar	0.511	10
Hinglish	0.586	4
Average	**0.718**	

6. Conclusion

We presented a lightweight yet effective approach to multilingual text detoxification, utilizing GPT-4o-mini in a prompt-based setting on the PAN 2025 ParaDetox benchmark. Our system does not require task-specific fine-tuning and relies solely on a single system instruction. It delivers strong performance in both initial and post-evaluation phases, ranking in the top 3 for 6 out of the 15 languages. The post-evaluation results further validated our model's robustness. We secured fourth place in the parallel and third in the non-parallel data tracks, achieving high J scores across English, German, Hebrew, and Spanish. Although our method struggled in low-resource and code-switched languages such as Hinglish and Tatar, it consistently performed well in morphologically diverse and high-resource languages. These results demonstrate the efficacy of prompt engineering in guiding multilingual LLMs for detoxification tasks and highlight the importance of future work on adaptation strategies for low-resource and complex linguistic settings.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] T. Alsubait, A. Alarifi, W. Alosaimi, The impact of online toxicity on social media platforms, in: 2021 International Conference on Computer and Information Sciences (ICCIS), IEEE, 2021, pp. 1–6.

[2] Z. J. Zhang, E. Sheng, E. Wallace, Detoxify: A robust pretrained transformer for toxic comment classification, arXiv preprint arXiv:2104.07367 (2021).

[3] L. Dos Santos, B. Silva, G. Rezende, A survey on text detoxification: Challenges, methods, and evaluation, in: Findings of the Association for Computational Linguistics: ACL 2023, 2023.

[4] 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 theSixteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.

[5] V. Protasov, PAN 2024 Multilingual TextDetox: Exploring Cross-lingual Transfer in Case of Large Language Models, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. Herrera (Eds.), Working Notes Papers of the CLEF 2024 Evaluation Labs, CEUR-WS.org, 2024, pp. 2852–2857. URL: http://ceur-ws.org/Vol-3740/paper-274.pdf.

[6] E. Spertus, Smokey: automatic recognition of hostile messages, in: Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence, AAAI'97/IAAI'97, AAAI Press, 1997, p. 1058–1065.

[7] D. Dementieva, D. Moskovskiy, D. Dale, A. Panchenko, Exploring methods for cross-lingual text style transfer: The case of text detoxification, 2023. URL: https://arxiv.org/abs/2311.13937. arXiv:2311.13937.

[8] M. Iglesias, Exploring toxic lexicon similarity methods with the DRG framework on the toxic style transfer task, Master's thesis, KTH, School of Electrical Engineering and Computer Science (EECS), 2023.

[9] D. Moskovskiy, S. Pletenev, A. Panchenko, LLMs to replace crowdsourcing for parallel data creation? the case of text detoxification, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational

Linguistics, Miami, Florida, USA, 2024, pp. 14361–14373. URL: https://aclanthology.org/2024.findings-emnlp.839/. doi:10.18653/v1/2024.findings-emnlp.839.

[10] 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.

[11] S. Pletenev, Memu_pro_kotow at PAN 2024 TextDetox: Uncensored Llama3 Helps to Censor Better, 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 Workshop Proceedings, CEUR-WS.org, 2024. PAN 2024 TextDetox workshop contribution.

[12] E. Rykov, K. Zaytsev, I. Anisimov, A. Voronin, Smurfcat at pan 2024 textdetox: Alignment of multilingual transformers for text detoxification, 2024. URL: https://arxiv.org/abs/2407.05449. arXiv:2407.05449.

[13] 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. Radev, A. F. Aji, K. Almubarak, S. Albanie, Z. Alyafeai, A. Webson, E. Raff, C. Raffel, Crosslingual generalization through multitask finetuning, in: A. Rogers, J. L. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 15991–16111. URL: https://doi.org/10.18653/v1/2023.acl-long.891. doi:10.18653/v1/2023.acl-long.891.

[14] J. Hong, N. Lee, J. Thorne, Orpo: Monolithic preference optimization without reference model, CoRR, abs/2403.07691, 2024. doi:10.48550/arXiv.2403.07691, arXiv:2403.07691, https://doi.org/10.48550/arXiv.2403.07691.

[15] V. Protasov, Pan 2024 multilingual textdetox: Exploring cross-lingual transfer in case of large language models, 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 Workshop Proceedings, CEUR-WS.org, 2024.

[16] N. Sushko, Pan 2024 multilingual textdetox: Exploring different regimes for synthetic data training for multilingual text detoxification, 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 Workshop Proceedings, CEUR-WS.org, 2024.

[17] 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. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

[18] 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: Generative ai detection, multilingual text detoxification, multi-author writing style analysis, and generative plagiarism detection, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp. 434–441.

A. Online Resources

The source code is available here.

- GitHub