Team cake at TextDetox CLEF 2025/Multilingual Text Detoxification 2025: A Multilingual Text Detoxification Method Based on Chain-of-Thoughts Prompting Approach

Notebook for PAN at CLEF 2025

Jiangao Peng[1], Kaiyin Sun[2], Kaichuan Lin[1], Zhankeng Liang[1] and Zhongyuan Han[1,*]

[1] *Foshan University, Foshan, China*
[2] *Foshan No.3 Middle School, Foshan, China*

Abstract

This paper presents a multilingual text detoxification method based on a chain-of-thoughts prompting approach for the PAN at CLEF 2025. Multilingual text detoxification aims to transform toxic texts into neutral versions while preserving the original meaning and grammatical structure. Our method leverages large language models (LLMs) to classify toxic sentences and detoxify them through carefully designed prompts. We evaluate our approach on the PAN 2025 multilingual text detoxification task, demonstrating its potential and stability in handling various languages.

Keywords

PAN 2025, Text Detoxification, Large Language Models, Chain-of-Thoughts Prompting, Text Classification

1. Introduction

Multilingual text detoxification is an important downstream task in natural language processing [1, 2, 3, 4, 5, 6, 7], aiming to transform toxic texts into neutral versions while preserving the original meaning and grammatical structure. This task is significant for purifying the online language environment. Among existing methods, the approach proposed by Dementieva et al. [8] has achieved SOTA. However, their method employs K-means clustering for text clustering, without leveraging large language models (LLMs). This limits the full utilization of the advantages that LLMs can offer. To address this issue, this paper proposes using LLMs to classify toxic sentences. We input toxic sentences along with their text features into the LLM to obtain the type of toxic sentences. This information then enables further detoxification. We evaluate our method on the PAN 2025 multilingual text detoxification task.

2. Method

Our method consists of three key stages: 1) Extracting toxic text features; 2) Classifying toxic text by features; 3) Text detoxification via text features and classification. All three steps are implemented through carefully designed prompts fed into the large language model.

2.1. Extracting Toxic Text Features

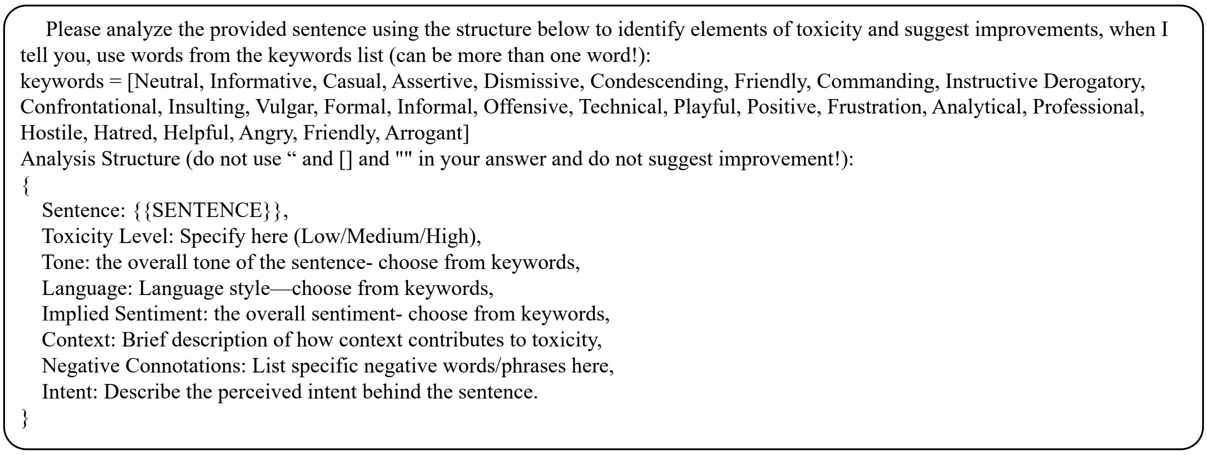
In this stage, we adopt a similar approach to that proposed by Dementieva et al. [8], leveraging the capabilities of LLMs to meticulously extract text features from the input sentences. The prompt provided to the LLM is designed to elicit a comprehensive analysis of the text, focusing on identifying elements

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

[*] Corresponding author.

wyd1n910@gmail.com (J. Peng); sunkaiyin123@163.com (K. Sun); 1441584221bruan@gmail.com (K. Lin); oooo1390329007@163.com (Z. Liang); hanzhongyuan@gmail.com (Z. Han)
0009-0006-3780-5023 (J. Peng); 0009-0001-7966-8390 (K. Sun); 0009-0001-7843-2888 (K. Lin); 0009-0009-6124-5494 (Z. Liang); 0000-0001-8960-9872 (Z. Han)

of toxicity and categorizing various aspects of the sentence structure and semantics. Figure 1 illustrates the structure of this prompt.



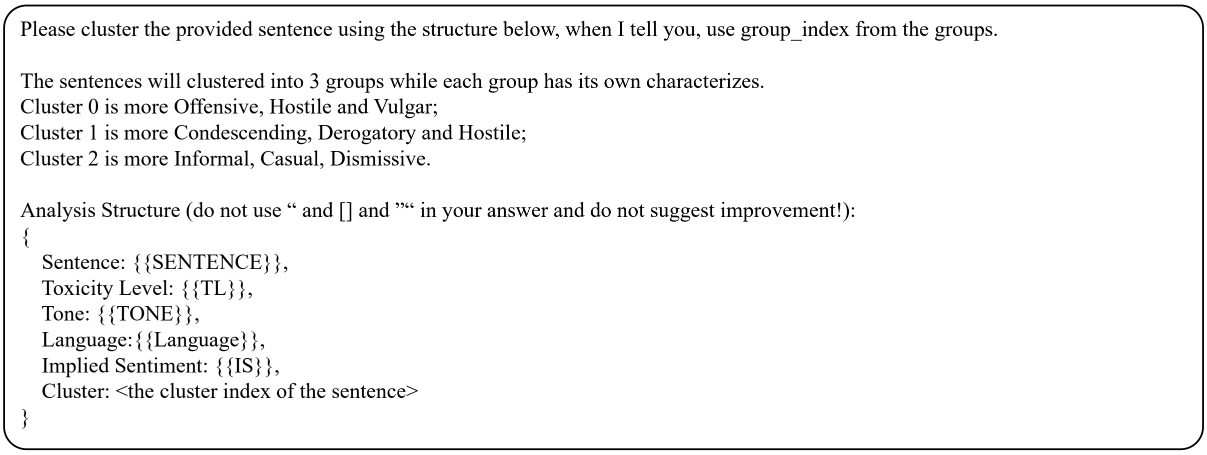
```
   Please analyze the provided sentence using the structure below to identify elements of toxicity and suggest improvements, when I
tell you, use words from the keywords list (can be more than one word!):
keywords = [Neutral, Informative, Casual, Assertive, Dismissive, Condescending, Friendly, Commanding, Instructive Derogatory,
Confrontational, Insulting, Vulgar, Formal, Informal, Offensive, Technical, Playful, Positive, Frustration, Analytical, Professional,
Hostile, Hatred, Helpful, Angry, Friendly, Arrogant]
Analysis Structure (do not use “ and [] and "" in your answer and do not suggest improvement!):
{
   Sentence: {{SENTENCE}},
   Toxicity Level: Specify here (Low/Medium/High),
   Tone: the overall tone of the sentence- choose from keywords,
   Language: Language style—choose from keywords,
   Implied Sentiment: the overall sentiment- choose from keywords,
   Context: Brief description of how context contributes to toxicity,
   Negative Connotations: List specific negative words/phrases here,
   Intent: Describe the perceived intent behind the sentence.
}
```

Figure 1: Structure of the Prompt for Extracting Toxic Text Features

By employing this structured prompt, the LLM systematically extracts these features and generates a detailed profile of the input text. This profile serves as the basis for the subsequent clustering and classification stage, where the extracted features are utilized to categorize the text into different clusters based on their toxicity profiles and other semantic characteristics.

2.2. Classifying Toxic Text by Features

Dementieva et al. [8] employed k-means algorithms for toxic text classification. However, these algorithms may struggle to capture complex semantic relationships and patterns when dealing with text clustering. In contrast, we directly utilize an LLM for classification by inputting the extracted text features into the model. With its powerful understanding and analytical capabilities, the LLM can uncover latent semantic connections between text features, thereby achieving more precise and semantically logical clustering of texts. Figure 2 illustrates the structure of this prompt.



```
Please cluster the provided sentence using the structure below, when I tell you, use group_index from the groups.

The sentences will clustered into 3 groups while each group has its own characterizes.
Cluster 0 is more Offensive, Hostile and Vulgar;
Cluster 1 is more Condescending, Derogatory and Hostile;
Cluster 2 is more Informal, Casual, Dismissive.

Analysis Structure (do not use “ and [] and ”“ in your answer and do not suggest improvement!):
{
   Sentence: {{SENTENCE}},
   Toxicity Level: {{TL}},
   Tone: {{TONE}},
   Language:{{Language}},
   Implied Sentiment: {{IS}},
   Cluster: <the cluster index of the sentence>
}
```

Figure 2: Structure of the Prompt for Classifying Toxic Text by Features

2.3. Text Detoxification via Text Features and Classification

To achieve text detoxification, we extended the approach of Dementieva et al. [8] by designing a prompt structure that leverages both text features and classification information. This approach allows us to systematically convert toxic sentences into neutral, informal, or casual ones while preserving their original meaning. By specifying the cluster characteristics and toxicity levels, the LLM can better understand the context and nuances of the input sentence, enabling it to generate an appropriate non-toxic version. Figure 3 illustrates the structure of this prompt.

```
Please detoxify the provided sentence using the structure below without changing the real meaning of the sentence.
The sentences are clustered into 3 groups while each group has its own characterizes.
Cluster 0 is more Offensive, Hostile and Vulgar;
Cluster 1 is more Condescending, Derogatory and Hostile;
Cluster 2 is more Informal, Casual, Dismissive.
For each sentence and cluster that I give, make the sentence non-toxic by making it Neutral/Informal/Casual without changing the
meaning.
Analysis Structure (do not use ” and [] and “” in your answer and do not suggest improvement!):
{
   Sentence: {{Sentence}},
   Toxicity level: {{TL}},
   Cluster: {{Cluster}},
   Fixed sentence: <the non-toxic sentence after making it Neutral/Informal/Casual without changing the meaning, use the same
language as the original sentence>;
}
```

Figure 3: Structure of the Prompt for Text Detoxification via Text Features and Classification

3. Experiment

3.1. Dataset

PAN 2025 multilingual text detoxification task has been improved and expanded in terms of data[1], evaluation metrics, and task settings. In this year's competition, six additional languages have been introduced, none of which provide parallel datasets.

3.2. Settings

In this competition, we utilized the DeepSeek-R1[9] provided by Volcengine[2] .

3.3. Evaluation

The official provided four metrics [3] . Each metric component lies in the range $[0, 1]$.

- **Style Transfer Accuracy (STA):** Classify its level of non-toxicity.
- **Content preservation (SIM):** Given two texts (original toxic sentence and generated paraphrase), evaluate the similarity of their content.
- **ChrF1:** To estimate the adequacy of the text and its similarity to the human-written detoxified references.
- **Joint (J):** To have the one common metric for leaderboard estimation, the official will compute $Joint$ metrics as the mean of $STA * SIM * FL$ per sample.

[1] https://huggingface.co/datasets/textdetox/multilingual_paradetox_test
[2] https://www.volcengine.com/
[3] https://codalab.lisn.upsaclay.fr/competitions/22396#learn_the_detailsevaluation

3.4. Baseline

- **golden annotation**: Human-written detoxified references.
- **baseline_gpt4**: A baseline model using GPT-4 for detoxification.
- **baseline_mt0**: A baseline model using the mT0[4] model for detoxification.
- **baseline_o3mini**: A baseline model using the o3-mini model for detoxification.
- **baseline_gpt4o**: A baseline model using a variant of GPT-4 for detoxification.
- **baseline_duplicate**: A simple baseline that duplicates the toxic input as the output.
- **baseline_backtranslation**: A baseline using backtranslation for detoxification. It translates the input to a language with a strong detoxification model, detoxifies it, and translates it back to the target language.

3.5. Result

The official provides an evaluation using the LLM-as-a-Judge approach. Specifically, they have fine-tuned the Llama-3.1-8B-Instruct [5] model on the manual annotations from the TextDetox2024 [6] dataset.

Table 1 shows the final results of the LLM-as-a-Judge[7] evaluation. The *golden annotation* model performed the best, with an average score of 0.828, especially achieving a high score of 0.904 in Japanese. The *Team cake (our method)* model ranked second with an average score of 0.674. Although this score is lower than that of *golden annotation*, it outperformed all other baseline models. This indicates that the *Team cake* model has considerable potential and stability in multilingual evaluation tasks, although there is still room for improvement in certain languages. Overall, the *golden annotation* model demonstrated very stable performance across all languages, while the *Team cake* model showed competitive strength comparable to the *golden annotation* in the majority of languages.

Table 1
LLM-as-a-Judge: Results for New Six Languages

Team/Model	Average	Italian	Japanese	Hebrew	French	Tatar	Hinglish
golden annotation	0.828	0.893	0.904	0.783	0.724	0.780	0.887
Team cake(our method)	0.674	0.791	0.796	0.581	0.853	0.436	0.584
baseline_gpt4	0.662	0.790	0.779	0.578	0.865	0.438	0.524
baseline_mt0	0.641	0.749	0.711	0.501	0.793	0.598	0.494
baseline_o3mini	0.559	0.748	0.661	0.497	0.826	0.209	0.411
baseline_gpt4o	0.526	0.697	0.680	0.370	0.718	0.327	0.363
baseline_duplicate	0.429	0.455	0.442	0.407	0.460	0.421	0.387
baseline_backtranslation	0.254	0.333	0.147	0.349	0.503	0.054	0.139

4. Conclusion

In this paper, we proposed a multilingual text detoxification method based on a chain-of-thoughts prompting approach. Our method effectively utilizes large language models to extract toxic text features, classify toxic sentences, and detoxify them while preserving their original meaning. The experimental results on the PAN 2025 multilingual text detoxification task show that our method has considerable potential and stability, outperforming other baseline models. Although there is still room for improvement in certain languages, our approach demonstrates its effectiveness in handling multilingual text detoxification tasks. Future work will focus on further enhancing the performance of our method and exploring more advanced techniques to improve text detoxification across different languages.

[4] https://huggingface.co/s-nlp/mt0-xl-detox-orpo
[5] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[6] https://github.com/textdetox/textdetox_clef_2024/tree/main/human_evaluation_results
[7] https://pan.webis.de/clef25/pan25-web/text-detoxification.html#results

5. Acknowledgments

This work is supported by the Social Science Foundation of Guangdong Province, China (No.GD24CZY02)

Declaration on Generative AI

During the preparation of this work, we used a Large Language Model; specifically, we employed DeepSeek-R1, provided by Volcengine, for the following purposes: extracting toxic text features, classifying toxic texts, and performing text detoxification. We also used basic AI-powered text-editing tools to check the grammar and spelling of the manuscript content. After using these tools, we reviewed and edited the content as needed and take full responsibility for the publication's content.

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. Dale, A. Voronov, D. Dementieva, V. Logacheva, O. Kozlova, N. Semenov, A. Panchenko, Text detoxification using large pre-trained neural models, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 7979–7996.

[3] 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.

[4] 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, in: Proceedings of the RUSSE-2022 Shared Task, 2022. doi:10.28995/2075-7182-2022-21-114-131.

[5] 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.

[6] D. Dementieva, D. Moskovskiy, D. Dale, A. Panchenko, Exploring methods for cross-lingual text style transfer: The case of text detoxification, in: Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 1083–1101.

[7] 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.

[8] D. Dementieva, N. Babakov, A. Ronen, A. A. Ayele, N. Rizwan, F. Schneider, X. Wang, S. M. Yimam, D. Moskovskiy, E. Stakovskii, et al., Multilingual and explainable text detoxification with parallel corpora, arXiv preprint arXiv:2412.11691 (2024).

[9] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al., Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, arXiv preprint arXiv:2501.12948 (2025).