A Multilingual Text Detoxification Method Based on Few-shot Learning and CO-STAR Framework

Notebook for PAN at CLEF 2024

Jiangao Peng, Zhongyuan Han*, Huan Zhang, Jingyan Ye, Chang Liu, Biao Liu, Mingcan Guo, Haoyang Chen, Zijie Lin and Yujiao Tang

Foshan University, Foshan, China

Abstract

Multilingual text detoxification is a natural language processing downstream task that inputs toxic sentences, and then outputs a neutral version that preserves the original meaning and grammar. To address this task, our method proposes a novel method that integrates the CO-STAR prompting framework with few-shot learning, aided by a chat model. In the final manual evaluation of PAN 2024, we achieved the highest scores in the Chinese and English categories, with 0.91 and 0.86, respectively.

Keywords

PAN 2024, Multilingual Text Detoxification, CO-STAR, Few-shot Learning, Chat Model

1. Introduction

Disclaimer. *Please be aware that you may come across offensive or toxic language in this paper due to its subject matter.*

Identification of toxicity in user texts is an active region. Researchers show that social media's strategy of limiting users' profanity causes people to adopt various countermeasures, such as misspellings, abbreviations, and the fast pace of profane slang evolution [1]. To reduce toxic language on the web more effectively, the researchers have proposed a more proactive method to transform toxic texts into neutral ones [2]. In this study, we propose a novel method that integrates the CO-STAR prompting framework with few-shot learning in a chat model to address the PAN at CLEF 2024 task [3]: Multilingual Text Detoxification. In this competition, we receive toxic sentences in multiple languages from all over the globe: English, Chinese, and the other seven languages, and then produce a neutral version of the text in the corresponding language [4]. We are only familiar with English and Chinese; therefore, we have focused on text detoxification for these two languages. The final manual evaluation results show our method's effectiveness in both languages, achieving the first place.

2. Related Work

Researchers propose a natural language processing downstream task called text detoxification to combat internet toxic information more proactively. In 2021, researchers presented two robust unsupervised methods [2] in English, CondBERT, and ParaGedi, which focus respectively on unsupervised toxic word replacement and unsupervised text paraphrase rewriting. They were the SOTA method of the time. Both methods were migrated to the Russian language and got good results [5]. In 2022, researchers introduce **ParaDetox** [6], a novel English parallel corpus collection for detoxification tasks, and only simply train

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
wyd1n910@gmail.com (J. Peng); hanzhongyuan@gmail.com (Z. Han*); 13553204558@163.com (H. Zhang); yjy20030604@gmail.com (J. Ye); lc965024004@gmail.com (C. Liu); hyticen@gmail.com (B. Liu); gmc9812@163.com (M. Guo); hoyo.chen.i@gmail.com (H. Chen); lamlovezz113@gmail.com (Z. Lin); 125375136@qq.com (Y. Tang)
0009-0006-3780-5023 (J. Peng); 0000-0001-8960-9872 (Z. Han*); 0009-0006-3869-5858 (H. Zhang); 0009-0001-4021-9942 (J. Ye); 0009-0000-0887-9273 (C. Liu); 0009-0000-3031-9758 (B. Liu); 0000-0002-4977-2138 (M. Guo); 0000-0003-3223-9086 (H. Chen); 0009-0009-1492-809X (Z. Lin); 0009-0009-3996-6324 (Y. Tang)

in Bart [7] with default configuration, which significantly boosts the performance of detoxification models compared to SOTA unsupervised methods. In 2022, the first Russian detoxification competition [8], using a newly created Russian parallel corpus and manual evaluation to assess the performance of various detoxification models, showed that under the condition of using a sizeable parallel database, the use of end-to-end language models can complete the text detoxification task well. Based on these works, we know that we can use parallel corpora to improve the performance of text detoxification.

The methods of GPT [9] Models is a novel line covering toxic and neutral sentences. In 2021, researchers showed that a small training dataset could improve the performances of GPT-2 [10] for detoxifying texts for the Russian language [5]. In 2023, researchers used GPT-3 [11] with ParaDetox for text detoxification [12]. To show the generalization capabilities of GPT, Researchers evaluated GPT-4 [13] performance on 25 diverse NLP tasks, finding only an average 25% quality loss compared to SOTA solutions [14]. Drawn by the detoxification and generalization capabilities of GPT, we have decided to utilize a ChatGPT-like model(Kimi [1]) to address the task.

Prompt provides an efficient way to leverage the power of pre-trained language models, enabling models to perform well even in few-shot scenarios [14]. In 2022, researchers explored prompt engineering can significantly improve the ability of large language models to perform complex reasoning [14]. In 2023, Sheila Teo [15] won in The GPT-4 Prompt Engineering Competition of Singapore by using the CO-STAR framework, which provides a structured approach to designing effective prompts, ensuring optimal responses from large language models. Those work inspired us to apply the prompt statement based on the CO-STAR architecture to to make the model outputs more consistent with our requirements.

3. Method

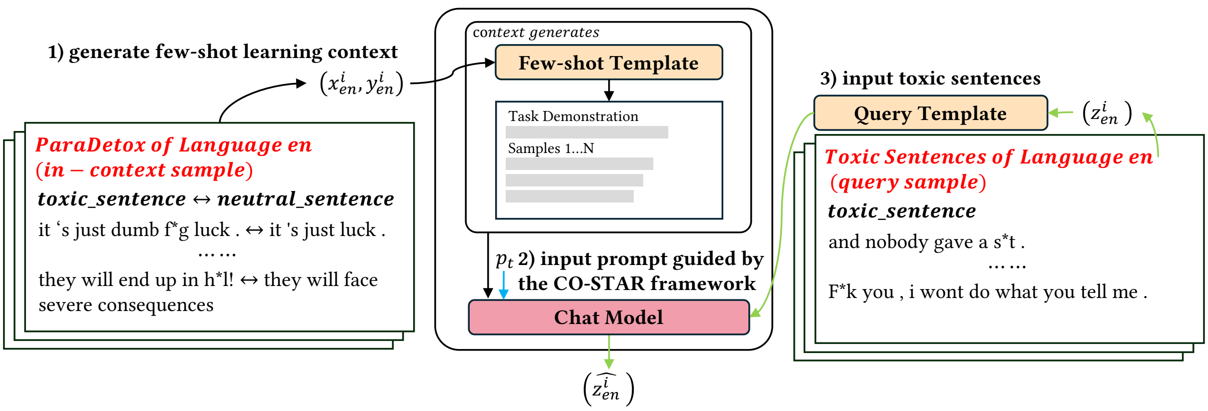


Figure 1: Architecture of the Detoxification Model

Our method involves three main steps: 1) generate a few-shot learning context, 2) input prompt guided by the CO-STAR framework, and 3) input toxic sentences. The official provides a multilingual parallel dataset and we use a few-shot learning context to make the chat model adapt to this task. We structure prompts using the CO-STAR framework that considers all the key aspects that influence the effectiveness and relevance of the model's response, leading to more optimal responses. Ultimately, we insert the target language's toxic sentences into the query template and then input them into the chat model to get the neutral version. We have selected Kimi [16] as our chat model and are adopting it without fine-tuning. Figure 1 shows the model architecture for completing text detoxification in English as an example, without losing generality for the other eight languages. Our method is detailed below:

[1]https://kimi.ai

3.1. Generate Few-shot Learning Context

This section shows how we generate the contents of a few-shot learning context.

Task Demonstration

现在介绍释义改写任务：
给定一段有毒的文本，以无毒的方式重写它，同时尽可能保留主要内容。旨在呈现句子的中立版本，保留有意义的内容。
处理方案：仅处理显性类型的毒性（明显存在淫秽和粗鲁的词汇，但仍然存在有意义的中立内容），而不处理隐性类型，例如讽刺、被动攻击性或对某些群体的直接仇恨找不到中性内容的地方。
下面列举了400个互联网上，**英语**的可能带冒犯文字改写为中性的话的改写的例子
(The paraphrase rewriting task is now introduced:
Given a toxic text, rewrite it non-toxic while retaining as much of the main content as possible. The aim is to present a neutral version of the sentence, retaining meaningful content.
Processing scheme: In our task, we will work only with explicit types of toxicity—the apparent presence of obscene and rude lexicon where there is still meaningful neutral content present—and do not work with implicit types—like sarcasm, passive aggressiveness, or direct hate to some group where no neutral content can be found.
Below is a list of 400 examples of rewrites on the Internet where potentially offensive text in **English** is rewritten as neutral words)

Samples 1...N: (x^i_{en}, y^i_{en}) with Few-shot Template

示例(*Sample*)*i*
$< |$toxic_sentence$| > x^i_{en}$
$< |neutral\_sentence| > y^i_{en}$

Figure 2: Detail of the Context Generates. Regardless of the language we are dealing with, we have used Chinese prompts, and the translations of the Chinese prompts are in parentheses.

Table 1
Sample English Pairs of the Dev Dataset

toxic_sentence	neutral_sentence	lang
it 's just dumb f*g luck .	it 's just luck .	en
they will end up in h*l!	they will face severe consequences .	en

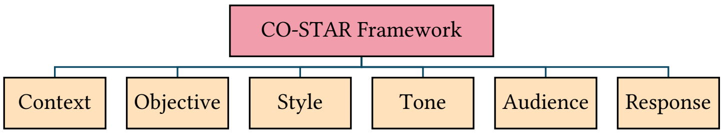
ParaDetox of Language en (in-context sample): During the test phase, the official presented parallel multilingual data, including English (en), Spanish (es), German (de), Chinese (zh), Arabic (ar), Hindi (hi), Ukrainian (uk), Russian (ru), and Amharic (am). This comprehensive dataset comprises 400 pairs for each of the nine languages, formatted as $< toxic\_sentence, neutral\_sentence, lang >$. Table 1 displays sample English pairs of the dev dataset for illustrative purposes.

Context Generates:

- ***Task Demonstration:*** To help the large chat model developed by the Chinese team understand the meaning of the text detoxification task, we carefully referenced the official English description of the task and then paraphrased it into the Chinese version. The top of Figure 2 shows the Chinese descriptor for the English text detoxification task.
- ***Few-shot template:*** To facilitate the model's understanding of the required neutral version of the toxic text, we provide a few-shot template that incorporates pairs of toxic and neutral sentences $(x^1_{en}, y^1_{en}), \ldots, (x^n_{en}, y^n_{en})$ all from the target language parallel dataset. Given that the chat modal we have adopted, Kimi, aligns with the Chinese language, we use Chinese keywords instead of English ones. The bottom of Figure 2 shows the template format. Ultimately, we insert N(here, $N = 400$) samples under the Task Demonstration.

Chat Model: We choose Kimi as our base chat model. Kimi is a powerful assistant developed by the Chinese company Moonshot AI, reportedly featuring around 20 billion parameters and supporting multiple languages.

3.2. Input Prompt Guided by The CO-STAR Framework



(a) CO-STAR Framework

CONTEXT
我现在正在完成自然语言处理的一项下游数据处理任务，我提供的*txt*文件是我准备的资料，我需要把一批$<|toxic_sentence|>$英文按照文件要求转换为$<|neutral_sentence|>$英文版本 (I am currently completing a downstream data processing task in natural language processing. The TXT file I provided contains the material I have prepared. I need to convert a batch of $<$ *toxic_sentence* $>$ **English** according to the requirements of the file into $<$ *neutral_sentence* $>$ **English** versions.)
OBJECTIVE
我会提供一批$<|toxic_sentence|>$英文给你，请你给我$<|neutral_sentence|>$英文版本，来完成任务 (I will provide you with a batch of $<$ toxic_sentence $>$ **English**; please give me the $<$ neutral_sentence $>$ **English** versions to complete the task.)
STYLE
你的风格应该像严谨的懂得文学的程序员，严格完成任务 (Your style should be like that of a rigorous programmer who understands literature and strictly completes the task.)
TONE
专业的 (Professional.)
AUDIENCE
懂得英文的成年评委，他们会判断你转换的$<|neutral_sentence|>$英文是否符合要求 (Adult judges who understand **English** will judge whether the $<$ *neutral_sentence* $>$ **English** you have converted meets the requirements.)
RESPONSE
返回如下 *json* 格式
[{*toxic_sentence*: "", *neutral_sentence*: ""}] (Return in the following JSON format:
[{"toxic_sentence": "", "neutral_sentence": ""}])

(b) Example for English Text Detoxification

Figure 3: Prompt Guided by the CO-STAR Framework. Regardless of the language we are dealing with, we have used Chinese prompts, and the translations of the Chinese prompts are in parentheses.

Practical cue construction is crucial to getting the best response from a large-scale language model (LLM). The CO-STAR framework, a brainchild of GovTech Singapore's Data Science & AI team [2], is a practical template for constructing cues. Figure 3a shows the CO-STAR framework. Here is how it works on this task. **(C) Context:** We inform the large model that we are currently undertaking a downstream data processing task in natural language processing and clarify that the TXT file we have provided contains the prepared materials and the AI need to convert a batch of $<|toxic\_sentence|>$ text into $<|neutral\_sentence|>$ versions according to the requirements specified in the file; **(O) Objective:** We define we will provide the model with a batch of $<|toxic\_sentence|>$ of one language and we would want the AI to return us a $<|neutral\_sentence|>$ version of them to complete the task; **(S) Style:** We specify want the LLM's style should be that of a diligent programmer who possesses a profound understanding of literature, while strictly adhering to the task requirements; **(T) Tone:** We set Professional as the attitude of the response; **(A) Audience:** We identify the understanding adult judges proficient in the target language as the intended recipients of the response, and **(R) Response:** We required a JSON output format like ***[toxic_sentence: "", neutral_sentence: ""]***.

3.3. Input Toxic Sentences

In this section, we show how we insert the target language's toxic sentences into the query template and then input them into the chat model to get the neutral version.

[2]https://www.tech.gov.sg/capability-centre-dsaid

Q: #CONTEXT# $[Upload\ File]$ ***en_kimi_prompt.txt***
A: $[Response]$
Q: $[Input\ prompt\ guided\ by\ the\ CO-STAR\ framework\ for\ English\ text\ detoxification]$
A: $[Response]$
Q: <|toxic_sentence|>[user_19]and nobody gave a s*t .
A: 以下是转换后的中性句子(Here are the neutral sentences after the transformation.)

```
[
  {
    "toxic_sentence": "[user_19]and nobody gave a s*t.",
    "neutral_sentence": "[user_19]and nobody cared."
  }
]
```

在转换过程中，我使用了更为中性的词汇来替换原文中的冒犯性语言，同时尽量保持了原句的意图和语境

（During the conversion process, I used more neutral vocabulary to replace the offensive language in the original text, while trying to maintain the original sentence's intention and context.）

Figure 4: The Real Detoxification Process and the Results of the English Samples in Table 2 . **Note: 1)** part of asterisk masked profanity is only in the paper, and the original data is used directly in the real case; and **2)** we only extract the content located in the json snippet in the response.

Toxic Sentences of Language (query sample): During the test phase, the official presented multilingual test data. Table 2 displays some English toxic sentences from the test dataset for illustrative purposes.

Table 2
English Toxic Sentences from the Test Dataset with Query Template Variants

#	toxic_sentence	Query Template
19	and nobody gave a s*t .	<|toxic_sentence|>[user_19]and nobody gave a s*t .
82	F*k you , i wont do what you tell me .	<|toxic_sentence|>[user_82]F*k you , i wont do what you tell me .

Query Template: As Figure 1 shows, we defined z_{en}^{i} as the detox sentence i of the English language and then inserted it into query template $<\ |toxic\_sentence|\ >\ [user_i]\ z_{en}^{i}$. Table 2 shows shows the result after insertion. It mocks $user\_x$ input toxic content on the Internet. We will send the content obtained after inserting the query template into the Kimi model in batches through the dialogue box. With the help of the previously uploaded context files and prompts based on the CO-SART framework, the Kimi model will return formatted neutral sentences $\left(\widehat{z_{en}^{i}}\right)$. Figure 4 demonstrates the real detoxification process and results of the English samples from Table 2. Additionally, you can access the Kimi link [3] to revisit our conversation and continue the dialogue to try the English detoxification process.

4. Experiment

4.1. Dataset

This year's text detoxification task is a multilingual effort aimed at transforming toxic sentences into a natural form while preserving the core message.

[3]https://kimi.moonshot.cn/share/cp6ivkecp7f7f0107lr0

- **Dev Set [4] :** It contains 400 pairs of $< toxic\_sentence, neutral\_sentence >$ samples for each of 9 languages.
- **Test Set [5] :** It contains 600 numbers of toxic sentences for each language formatted as $< toxic\_sentence, lang >$.

According to the official information, for each language, the test and dev datasets are derived from the same set of $1k$ parallel pairs.

4.2. Settings

We repeated the following steps for all nine languages:

1. **Generate and Input a Few-shot Learning Context:** We used all the 400 samples of each language in the dev set to generate a few-shot learning contextm and then input to the Model, as we said in section 3.1. For different languages we just replace the part of the context that identifies the language. The traditional in-context method sends the entire context, along with one's questions, followed by a suffix, such as $>>>$, directly to the large model through a dialogue box. Kimi offers a file upload method, where we can first put the context content into a document in TXT format and then upload it to Kimi. In practice, we have found that the latter method is better for our task this time. We have yet to conduct an actual analysis of this. We speculate that the file upload method is like a knowledge-based method, where the large model first deeply understands the relevant knowledge before answering our questions. The traditional method, on the other hand, is more like everyday communication between two people.
2. **Input Prompt Guided by the CO-STAR Framework:** When dealing with different languages, we change the keyword identifying the language in the prompt statement to the appropriate language Figure 3b provides a detailed demonstration of the practical application of CO-STAR in this task, which would go directly to Kimi through a dialogue box.
3. **Input Toxic Sentences:** We directly insert the toxic sentences of the test set into the query template as described in section 3.3 and then input it to Kimi.
4. **Processing of Acquired Results:** Figure 4 realistically shows what the model returns to us. We would take the $neutral\_sentence$ field of the $JSON$ data returned by the model.

4.3. Evaluation

The official provided four metrics [6] . 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.

4.4. Baseline

The official provided four baselines for manual evaluation.

- **Delete:** Elimination of toxic keywords based on a predefined dictionary [7] for each language.

[4] https://huggingface.co/datasets/textdetox/multilingual_paradetox
[5] https://huggingface.co/datasets/textdetox/multilingual_paradetox_test
[6] https://codalab.lisn.upsaclay.fr/competitions/18243#learn_the_detailsevaluation
[7] https://huggingface.co/datasets/textdetox/multilingual_toxic_lexicon

- **Backtranslation:** A more sophisticated cross-lingual transfer method. Translate the input to English with NLLB-600M model[8], perform detoxification with English bart-base-detox model [9] , and translate back to the target language.
- **mt5** [10] : A supervised baseline which used **mt5-xl [17]** fine-tuned on the parallel dev set.
- **Human references:** Humans write them.

4.5. Result

Our method conducted text detoxification for all nine languages in the competition and submitted the results. Even if powerful models exist to classify texts and embed their meanings, human judgment is still the best for the final decision [8].

Table 3
Manual Evaluation Leaderboard. The results of the Manual evaluation with crowdsourcing on a random subsample of 100 texts per language. This leaderboard is based on the J scores of English(en). We highlight top1 results per each column. Our team name is **Team cake**.

User	en*	zh	es	de	ar	hi	uk	ru	am	average
Team cake	**0.91**	**0.84**	0.77	0.77	0.44	0.34	0.50	0.71	0.61	0.65
mkrisnai	0.89	0.34	0.83	0.70	0.63	0.73	0.73	0.78	0.49	0.68
Human References	0.88	0.93	0.79	0.71	0.82	0.97	0.90	0.80	0.85	0.85
SomethingAwful	0.86	0.53	**0.83**	**0.89**	0.74	0.86	0.69	**0.84**	0.71	**0.77**
bmmikheev	0.84	0.60	0.76	0.78	0.69	0.78	0.63	0.51	0.56	0.69
adugeen	0.83	0.60	0.73	0.70	0.82	0.68	**0.84**	0.76	0.71	0.74
ZhongyuLuo	0.73	0.56	0.52	0.01	0.49	0.49	0.42	0.68	**0.72**	0.51
backtranslation	0.73	0.34	0.56	0.34	0.42	0.33	0.23	0.22	0.54	0.41
nikita.sushko	0.70	0.47	0.62	0.79	**0.89**	0.84	0.67	0.74	0.68	0.71
VitalyProtasov	0.69	0.49	0.81	0.77	0.79	**0.87**	0.67	0.73	0.68	0.72
mT5	0.68	0.43	0.47	0.64	0.63	0.60	0.42	0.40	0.61	0.54
delete	0.47	0.43	0.55	0.57	0.65	0.65	0.60	0.49	0.63	0.56

Table 3 shows the leaderboard which we built it with all four official baselines as well as other teams that were #1 in each language from *Test Phase: Manual Evaluation Final Results* [11]. Several vital insights emerge when examining the performance across different languages. Firstly, we achieved the highest score of 0.91 in the English (en) language and 0.84 in Chinese (zh), respectively, surpassing all other users, which indicates a significant strength in English and Chinese processing capabilities. Compared to Human References, which represent the human benchmark, our method exceeds human performance in English, a noteworthy achievement. Secondly, we also demonstrate relatively good performance in German (de) and Russian (ru), scoring 0.77 and 0.71 respectively. This suggests that our method can effectively address cross-language detoxification tasks. However, we are facing challenges in Arabic (ar) and Hindi (hi), achieving scores of only 0.44 and 0.34 respectively. These low scores indicate that our model still requires improvement when dealing with these languages.

Table 4 shows examples of baseline models and our model output in English. *Delete* directly deletes profanity words based on the multilingual toxic lexicon database, resulting in the returned results being usually not fluent, and this method cannot handle profanity words that do not exist in the database. Although *mt5* can solve the problem of fluency, it cannot handle general toxic texts. *Backtranslation* still required a larger corpus to address the issue of generated text mismatching its original meaning. Our model, like mt5, uses a small dataset but generates fluent content with strong generalization ability.

[8]https://huggingface.co/facebook/nllb-200-distilled-600M
[9]https://huggingface.co/s-nlp/bart-base-detox
[10]https://huggingface.co/textdetox/mt5-xl-detox-baseline
[11]https://pan.webis.de/clef24/pan24-web/text-detoxification.html#results

Table 4
Examples of detoxifications by baseline models and our model. Inspired by the work of Daryna et al. [6] , we have indicated bad answers in red and good answers **in bold**.

Original	i did not vandalize anything you d*ck .	Yeah, you can f*ck off incel.	Be careful who you call a m*ron.
delete	i did not vandalize anything you .	Yeah, you can off incel.	Be careful who you call a m*ron.
mT5	I did not vandalize anything you did not do.	Yeah, you can *ck off incel.	Be careful who you call a m*ron.
backtranslation	**I did not vandalize anything.**	**Yeah, you can go away incel.**	Be careful who you call.
Team cake	**I did not vandalize anything.**	**Yeah, you can leave now.**	**Be careful who you insult.**

5. Conclusion

This study proposes a novel multilingual text detoxification method that combines the CO-STAR framework and few-shot learning, which is implemented by Kimi, a chat model. Our aim is to efficiently convert toxic texts in multilingual languages into neutral versions while preserving the original meaning and grammatical structure. We are only familiar with English and Chinese; therefore, we have focused on text detoxification for these two languages. In the PAN at CLEF 2024 competition, we achieved the highest scores in the Chinese and English categories, with 0.91 and 0.86, respectively, at manual evaluation, demonstrating a significant advantage in processing power for both languages.

The proposed method outperforms human benchmarks on English, exhibiting high efficiency and accuracy in text detoxification tasks. This is a noteworthy achievement in the field of natural language processing. While the results on English and Chinese are remarkable, the performance on other languages such as Arabic and Hindi is not as strong, indicating that the model still requires further optimization and improvement for these languages. We need to do more ablation experiments to study the effect of different prompt sentences, different chat models and other factors on the detoxification effect. Although the current prompt is based on the CO-SART framework, it is still written by human beings, and in the future we will study the automatic generation of optimal prompt statements. We also plan to employ large models to evaluate the detoxification effectiveness.

Acknowledgments

This work is supported by the National Social Science Foundation of China (22BTQ101)

References

[1] S. Sood, J. Antin, E. Churchill, Profanity use in online communities, in: Proceedings of the SIGCHI conference on human factors in computing systems, 2012, pp. 1481–1490.

[2] 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).

[3] 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.

[4] 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.

[5] 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.

[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] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461 (2019).

[8] 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.

[9] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018).

[10] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI (2019).

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[12] A. Pesaranghader, N. Verma, M. Bharadwaj, Gpt-detox: An in-context learning-based paraphraser for text detoxification, in: 2023 International Conference on Machine Learning and Applications (ICMLA), IEEE, 2023, pp. 1528–1534.

[13] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[14] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Computing Surveys 55 (2023) 1–35.

[15] S. Teo, How i won singapore's gpt-4 prompt engineering competition: A deep dive into the strategies i learned for harnessing the power of large language models (llms), Towards Data Science (2023). URL: https://towardsdatascience.com/how-i-won-singapores-gpt-4-prompt-engineering-competition-34c195a93d41, accessed: July 10, 2024.

[16] J. Chen, S. Li, Q. Huang, S. Yan, Z. Xie, Y. Lu, Application of kimi intelligent assistant in the teaching of water pollution control engineering course, International Journal of Education and Humanities 13 (2024) 39–43.

[17] 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: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 483–498. URL: https://aclanthology.org/2021.naacl-main.41.