SINAI at PAN 2024 TextDetox: Application of Chain of Thought with Self-Consistency Strategy in Large Language Models for Multilingual Text Detoxification

Notebook for PAN at CLEF 2024

María Estrella Vallecillo-Rodríguez[1], Arturo Montejo-Ráez[1] and María Teresa Martín-Valdivia[1]

[1]Computer Science Department, SINAI, CEATIC, Universidad de Jaén, 23071, Spain

Abstract

This article describes the participation of the SINAI research group in the shared task TextDetox (Multilingual Text Detoxification) in CLEF 2024. TThe proposed system for multilingual text detoxification employs Large Language Models (LLMs) utilizing a Self-Consistent Chain of Thought (CoT-SC) prompting strategy. This CoT-SC strategy consists of identifying the language of the toxic comment and then generating three different detoxified text proposals, the first proposal consists of removing the toxic words, the second of replacing the toxic words with neutral words, and the last of rewriting the toxic text in a neutral way. Subsequently, the selected LLM has to evaluate each generated neutral text according to the competition metrics. Finally, the model selects the best neutral text generated. Specifically with this proposal, we aim to evaluate the capacity of auto-evaluation and reasoning of LLM in different languages, including those with low resources. Our proposal was ranked 23rd in the automatic evaluation metrics and 11th in the final ranking with the manual evaluation.

Keywords

Multilingual Detoxification, Large Language Models, Chain of Thought with Self-Consistency, Text Generation,

1. Introduction

Social networks have allowed us all to be connected and know what is happening on the other side of the world in a few seconds. However, the inappropriate use of these social networks and the anonymity that these platforms allow make it easier to offend other users, and the network is filled with inappropriate comments such as toxic comments. The task organizers define toxic comments as those comments that contain obscene and rude language mixed with neutral content (explicit toxicity) and those comments that do not contain neutral text and are loaded with sarcasm, passive aggressiveness, or direct hatred towards some group or individual. Although in different research works, the term toxicity can have different definitions according to the aspects of toxic language they address [1] and have also been used to describe hate speech [2, 3], abusive [4], aggressive [5], and offensive language [6].

Due to all of the problems mentioned previously and thanks to the capacity of the new large language models to generate text or what is currently known as generative AI, it has been possible to explore different proactive strategies to mitigate offensive language in online environments, such as the automatic generation of counter-narratives [7, 8] or the strategy of text detoxification. In this case, the organizers centered at text detoxification and proposed the shared task TextDetox, where the main objective is to generate neutral alternatives to toxic comments. To do so, they focus on texts with explicit toxicity because of the complexity of detoxifying texts with implicit toxicity in which the initial intention of the comment is already toxic. To detoxify texts it is important to maintain as much

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

*Corresponding author.
†These authors contributed equally.
mevallec@ujaen.es (M. E. Vallecillo-Rodríguez); amontejo@ujaen.es (A. Montejo-Ráez); maite@ujaen.es (M. T. Martín-Valdivia)
0000-0001-7140-6268 (M. E. Vallecillo-Rodríguez); 0000-0002-8643-2714 (A. Montejo-Ráez); 0000-0002-2874-0401 (M. T. Martín-Valdivia)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

CEUR-WS.org/Vol-3740/paper-285.pdf

content as possible and the same writing style used by the author of the commentary. Furthermore, the authors propose to carry out this task for 9 languages without having sample detoxified data for all these languages in order to generate unsupervised and multilingual systems. The ultimate goal with which this task was proposed is to try to make the comments posted on social networks less toxic and therefore, these environments stop being so toxic.

In this paper, we present the systems developed as part of SINAI team participation in the TextDetox (Multilingual Text Detoxification) shared task [9] organized in the PAN Lab [10], at CLEF 2024. The aim of TextDetox is the creation of detoxification systems for 9 languages such as English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian, and Amharic with unsupervised and cross-lingual detoxification systems. For this reason, our proposal is the use of Large Language Models (LLMs) such as GPT-3.5 and Mistral with a Self-Consistency Chain of Thought (CoT-SC) prompt strategy [11]. With this strategy, we expect to generate good neutral texts to supply the toxic ones exploring different ways to solve the task using reasoning and argument by the models, In addition, with CoT-SC we can explore the models' ability to self-assess their responses.

CoT-SC is similar to Chain of Thought (CoT) [12] and surge because with CoT strategy in complex problems, if the initial decisions derail a solution, the approach will fall short. Due to this reason, we need a prompting approach that performs more deliberate planning and exploration when solving problems, like CoT-SC. With regard to the task of detoxification of texts, there is previous research using different prompting techniques, such as [13] in which a framework is created that takes the toxic input. Later the detoxification model generates the explanation of why the input is toxic, as well as a non-toxic version. The paraphrase detector will analyze the semantic similarity of the toxic and non-toxic pair and generate a warning if the pair is not semantically equivalent. To implement such a system they collect toxic data from several social networks, then generate with GPT texts of the opposite kind, filter the old data, and generate again with GPT on the one hand an explanation of why the texts are toxic and on the other hand the paraphrase of these texts. Two LLaMA models are then trained, one for explaining why a text is toxic and one for paraphrasing. During the whole process of corpus generation, it is very important to use a good prompt and specifically in this work LLaMA has been trained to be able to argue with CoT. Other works like [14], use Generative Pretrained Models (GPT) and different prompting techniques such as Few-Shot Learning (FSL) [15] or Zero-Shot Learning (ZSL) [16] to generate the detoxified sentence.

The rest of the paper is structured as follows: In Section 2 we describe the main strategy used to develop the system for the shared task. The data and the experimental methodology are described in Section 3. The results from the automatic and manual evaluation are presented in Section 4. Finally, we conclude with a discussion in Section 5.

2. System overview

The system developed for the TextDetox shared task at CLEF 2024 is described in this section.

We propose the use of LLMs with a prompt strategy called Chain of Thought with Self-Consistency.

This technique is similar to the Chain of Thought strategy, where the model resolves complex tasks by reasoning through a specific path. However, CoT has a limitation: the LLMs follow only one chain of thought and do not explore other potential reasoning approaches. With the CoT-SC strategy, we use diverse reasoning paths through few-shot CoT thoughts and select the most consistent answer from the generated options. This approach enables an LLM to self-evaluate its progress through intermediate thoughts, engaging in a deliberate reasoning process. Our aim in applying this strategy is (1) to explore different methods for generating neutral texts, (2) to study the model's capability to coherently evaluate each alternative, and (3) to analyze if LLMs can follow their thought process to select the best option and achieve good results.

The architecture of the prompt used in our task can be seen in Figure 1. Due to the fact that we have multilingual data, the first step of our prompt consists of identifying the language of the toxic text. Later, we specify to the LLM that it has to apply 3 different methods to detoxify the text, first, it has to

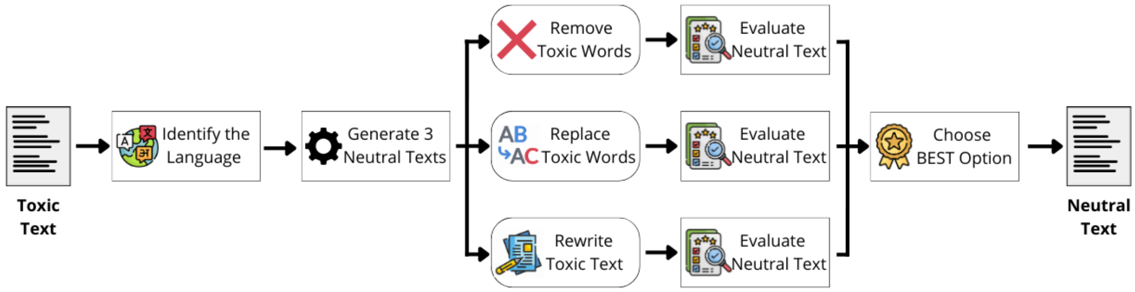

Figure 1: Proposed system for the TextDetox task. This system receives a toxic text and follows different steps such as identifying the language of the toxic text, generating three different neutral options, evaluating them, and choosing the final neutral text to give it as the output. In bold are marked the input and the output of our system.

remove the toxic words, in the second method it has to replace toxic words with neutral words and the last strategy consists of rewrite toxic texts in a neutral form. In the next step the model has to evaluate each option of detoxified text with the following questions: (1) 'Does the neutral text contain offenses or swear words?', (2) 'Do the neutral text and the TOXIC TEXT mean the same?', and (3) 'Is the neutral text grammatical?'. Each question refers to an evaluation metric of the shared task (explained in Section 4), first for *Style Transfer Accuracy*, the second for the *Content preservation*, and the last for the *Fluency*. Finally, with the evaluation of each neutral text, the model has to choose the best option, non-toxic, with the most content as the toxic words and without grammatical errors. The prompt used to generate the neutral texts is shown in Table 1.

Table 1
Prompt used to generate the neutral text in TextDetox shared task

Prompt
You are going to receive a TOXIC TEXT in a language. Your task consists of the following steps: 1. Identify which language the TOXIC TEXT is in. 2. Provide 3 different neutral alternatives in the identified language. The first one removing the toxic words, the second one replacing the toxic words with neutral words, and the last one rewritting the TOXIC TEXT in a neutral form. 3. Evaluate each alternative according to the following questions: 'Does the neutral text contain offenses or swear words?', 'Do the neutral text and the TOXIC TEXT mean the same?', and 'Is the neutral text grammatical?' 4. Based on this evaluation select the best neutral alternative that changes the minimum number of words of the TOXIC TEXT 5. Write the neutral text at the end of your answer with the keyword 'BEST PROPOSAL:'. TOXIC TEXT: {}

3. Experimental setup

3.1. Data

To run our experiments, we use the dataset provided by the organizers called "multilingual paradetox". This dataset contains 1000 pairs of toxic comments and their associated neutral text for each language contained in the dataset. Moreover the dataset includes toxic comments for 9 languages, that are English, Russian, Ukrainian, Spanish, German, Amhairc, Arabic, Hindi, and Chinese. To create the dataset, the organizers collected original toxic sentences from datasets previously used in other work or shared tasks. This dataset is splitted in two part, a dev set that contains 400 toxic comments for each language and the rest 600 comment for the test split.

Moreover, in the first epoch of the competition, the organizers provided English and Russian parallel corpora of several thousand toxic-detoxified pairs that participants could use to train the models. For other languages, no such corpora will be provided because the main challenge of this competition will be to perform unsupervised and cross-lingual detoxification.

Since we propose a system based on the CoT-SC strategy that involves the previous knowledge of the selected models, we use the English and Russian parallel corpora to make previous experiments with the prompt to use. We also use the multilingual paradetox dev set examples for this purpose.

3.2. Experiments and Selected Models

To achieve the goal of the TextDetox shared task, we propose the use of Chain of Thought with Self-Consistency strategy as is explained in Section 2. To apply the CoT-SC prompt strategy we selected two models, GPT-3.5 [17] from OpenAI that have 175B of parameters and knowledge of different languages. With this model, we think that we will achieve good results in the task due to this knowledge. The reason we decided to use GPT-3.5 instead of GPT-4 is because GPT-3.5 takes less time to generate a response and we consider that the knowledge that this model has about the task to be performed is more than sufficient to address it. Additionally, because GPT-3.5 is a private model and we don't have full control over its parameters, we decided to use Mistral-7B-instruct-v0.2 [18] an open LLM with less number of parameters.

Initially, as we began to look for the best prompt to use for the task, we also experimented with LLaMA2-13B-chat [19] but we decided not to continue using that model because it generated very long answers and was not concise enough to give the answer.

To achieve the goal of the TextDetox task, we established 500 as the limit of the max new tokens generation, and a seed of 42 to make the results replicable.

4. Results

In this section, we present the results obtained by the system developed as part of our participation in TextDetox 2024. The organizers use TIRA.io platform [20] to present the evaluation results. The results are provided with two types of evaluation, the automatic evaluation (Section 4.1), where some metrics are calculated using a gold standard is used and the manual evaluation (Section 4.2) where humans read the detoxified texts and assign a punctuation about whether the generated text by our model is good.

4.1. Automatic evaluation

The automatic evaluation is realized using the official metrics of the shared task. These metrics are:

- **Style Transfer Accuracy (STA):** This metric calculates the level of non-toxicity of the generated text. For this purpose a fine-tuned xlm-roberta-large [21] for toxicity binary classification is used.
- **Content preservation (SIM):** It evaluates the similarity of the content between the toxic text and the neutral one generated in our experiments. It is calculated as cosine similarity using LaBSe embeddings [22].
- **Fluency (FL):** This metric measures the adequacy of the neutral text, writing it without errors and similar to the human detoxified reference. To calculate the chrF measure [23] is used.

All of these metrics are in the range between 0 and 1. To have only a reference metric, a *Joint* metric is calculated as the mean of the result of STA*SIM*FL per instance of the dataset.

Table 2 shows the results obtained by our experiments in the Joint metric for each language. In this case, the results obtained by GPT-3.5 are divided between the results obtained using each option of the CoT-SC prompt (Removing toxic words, replacing toxic words with neutral ones, and rewriting the toxic text), all of these options are the same in Mistral model but this model does not identify in all of the cases the format to give the response and sometimes generate larger sequences achieving the limit of generation tokens without reasoning what is the best generated neutral text. When it is impossible

Table 2
Results of our experiments in the official metric of the competition for the test set of multilingual paradetox dataset. The submitted result is in bold

Experiment	Strategy	Average total	am	ar	de	en	es	hi	ru	uk	zh
GPT-3.5	Removing	0.350	0.133	0.340	0.445	0.478	0.433	0.233	0.443	0.507	0.140
	Replacing	0.324	0.113	0.317	0.441	0.416	0.410	0.208	0.413	0.455	0.140
	Rewritting	0.224	0.069	0.197	0.304	0.289	0.291	0.188	0.294	0.304	0.081
	CoT-SC	**0.309**	**0.097**	**0.283**	**0.403**	**0.413**	**0.404**	**0.225**	**0.397**	**0.436**	**0.126**
Mistral	CoT-SC	0.121	0.206	0.119	0.216	0.219	0.152	0.033	0.051	0.060	0.032

to differentiate the best option or each option in the text, we use the toxic comment as neutral. This occurs in a large number of cases, more than 1500, so in this table we will only show the results of applying CoT-SC. More information related to this type of problem is in Section 4.3.

Analyzing the results obtained by our models (Table 2), we can see that GPT-3.5 generally performs better than Mistral. Specifically, in the CoT-SC strategy it outperforms Mistral by far (0.309 to 0.121). On the other hand, if we focus on each neutral option that has been generated, we see that in the average metric what works the best is the option of eliminating toxic words, reaching 0.35 in the *Joint* metric. this is because it generates sentences in which the neutral sentence is identical to the neutral one and therefore maintains more content and eliminates toxicity. However, as we will see in the error section, it generates ungrammatical texts. This is mainly caused by the fact that it generates sentences in which the neutral sentence is identical to the neutral one and therefore maintains more content and eliminates toxicity. However, as we will see in Section 4.3, it generates ungrammatical texts. The option of replacing toxic words with neutral words works somewhat worse than the option of eliminating toxic words. This may be because the sentences change their content somewhat and are not as similar to the ones used for reference. The option of replacing toxic words with neutral words works somewhat worse than the option of eliminating toxic words. This may be because the sentences change their content somewhat and are not as similar to the ones used for reference. Finally, the option based on rewriting the toxic text is the one that obtains the worst results in these metrics as it is the one with the most variations concerning the toxic comment and probably the one that is furthest away from the reference ones used to calculate the fluency. As we can see all the options have their advantages and disadvantages, so we decided to send the option selected by the GPT-3.5 model when applying CoT-SC, since the detoxified sentence will perform as well as possible in terms of grammatical errors and toxicity removal, even if it is not always very similar to the one taken from the reference dataset.

To conclude, from this table, we can also analyze which languages the models are best adapted to and in which we need to invest more resources. As expected, Amharic is the one that obtains the worst results, as it is a language poor in resources to train this type of model so that they can learn from it. Looking at the language to which the models are best adapted for this task, there is one surprising fact: in the case of GPT-3.5, the language that obtains the best results is Ukrainian, followed by English, even though this model has been trained mostly with data in English. Mistral, on the other hand, does achieve the best performance in English.

4.2. Manual evaluation

For manual evaluation, 100 texts of each language are selected in a random form and evaluated by human crowdsourcing. For manual evaluation, participants can only submit one result, so we decided on the GPT-3.5 CoT-SC and achieved the results presented in Table 3. As we can see, when evaluating our model manually, the average metric increases, reaching a value of 0.57. This indicates the importance of manual and human annotator reviews of the texts that are generated because although the automatic metrics give you an idea of the performance of the system, they do not assimilate to the validity and real quality of the texts. Furthermore, we find that our system performs best for English followed by Arabic. This is curious, because Arabic is not a very common language when generating resources to train this kind of models. And it shows lower performance on complex languages for which it has received less training data such as Amharic, Chinese, or Hindi.

Table 3
Rank and results of the manual evaluation in test split of the multilingual paradetox dataset. Our team's submitted result is in bold.

Rank	Experiment	Average total	en	es	de	zh	ar	hi	uk	ru	am
0	Human References	0.85	0.88	0.79	0.71	0.93	0.82	0.97	0.90	0.80	0.85
1	SomethingAwful	0.77	0.86	0.83	0.89	0.53	0.74	0.86	0.69	0.84	0.71
...											
10	Yekaterina29	0.64	0.75	0.63	0.74	0.30	0.70	0.66	0.65	0.70	0.60
11	**SINAI**	**0.57**	**0.85**	**0.68**	**0.52**	**0.33**	**0.76**	**0.54**	**0.65**	**0.67**	**0.14**
12	gleb.shnshn	0.56	0.74	0.68	0.55	0.41	0.54	0.65	0.44	0.61	0.47
13	delete Baseline	0.56	0.47	0.55	0.57	0.43	0.65	0.65	0.60	0.49	0.63
14	mT5 Baseline	0.54	0.68	0.47	0.64	0.43	0.63	0.60	0.42	0.40	0.61
...											
18	gangopsa	0.50	0.74	0.20	0.72	0.37	0.61	0.75	0.48	0.00	0.61
19	backtranslation Baseline	0.41	0.73	0.56	0.34	0.34	0.42	0.33	0.23	0.22	0.54

Compared with the rest of the participants, we have reached an average of 11th place, which indicates that our system has a very similar performance to the rest of the participants. Moreover, our system has a very similar performance to the human benchmark for English (0.85 to 0.88 achieved by the Human References) and Spanish although we have room for improvement (0.68 to 0.83 achieved by SomethingAwful). For the rest of the languages, we should still invest efforts in obtaining better performances for other languages, as our performance is clearly inferior to the teams that obtain the highest scores such as Amharic (0.14 to 0.71 achieved by SomethingAwful) or Chinese in which we do not exceed the baseline applied by the backtranslation.

Table 4
Errors finds in the generation of detoxified text using GPT-3.5 model an the dev set of the TextDetox Dataset.

Id	Toxic Text	Response
1	Ah, porque se va y viene la hija puta. Están haciendo el subnormal apagando y encendiendo *(Ah, because the motherfucker comes and goes. They're doing the subnormal switching off and on.)*	1. This TOXIC TEXT is in Spanish. 2. Neutral alternatives: - Removing toxic words: Ah, porque se va y viene la. Están haciendo el subnormal apagando y encendiendo.*(Ah, because it comes and goes. They are doing the subnormal switching off and on.)* - Replacing toxic words: Ah, porque se va y viene la mujer. Están haciendo el tonto apagando y encendiendo.*(Ah, because the woman comes and goes. They're fooling around turning off and on)* - Neutral rewrite: Ah, ¿por qué se va y viene tan seguido? Están actuando de manera absurda apagando y encendiendo.*(Ah, why does it come and go so often? They are acting absurdly by switching off and on.)* 3. Evaluation: - Removing toxic words: It does not contain offenses or swear words. The meaning is retained, but the text is somewhat ungrammatical without the second word. - Replacing toxic words: It does not contain offenses or swear words. The meaning is slightly altered with the replacement of words. The text is grammatical. - Neutral rewrite: It does not contain offenses or swear words. The meaning is expressed differently but conveys a similar message. The text is grammatical. 4. Based on evaluation, the best neutral alternative is the neutral rewrite as it conveys a similar message, avoids toxic language, and maintains grammatical correctness. 5. BEST PROPOSAL: Ah, ¿por qué se va y viene tan seguido? Están actuando de manera absurda apagando y encendiendo. *(Ah, why does it come and go so often? They are acting absurdly by switching off and on.)*
2	¿Y este subnormal quién es? *(And who is this subnormal?)*	1. The TOXIC TEXT is in Spanish. Neutral alternatives: 1. Neutral alternative removing toxic words: ¿Y esta persona quién es? *(And who is this person?)* 2. Neutral alternative replacing toxic words: ¿Y este desconocido quién es? *(And who is this stranger?)* 3. Neutral alternative rewriting in a neutral form: ¿Quién es este individuo? *(Who is this individual?)* Evaluation: - Alternative 1: The neutral text does not contain offenses or swear words. It maintains the meaning of the TOXIC TEXT and is grammatical. - Alternative 2: The neutral text does not contain offenses or swear words. It changes the original words to a more neutral form but still carries the same meaning. It is grammatical. - Alternative 3: The neutral text does not contain offenses or swear words. It completely changes the wording but conveys a similar neutral meaning. It is grammatical. Based on the evaluation, the best neutral alternative that changes the minimum number of words of the TOXIC TEXT is Alternative 1. BEST PROPOSAL: ¿Y esta persona quién es?*(And who is this person?)*

4.3. Error Analysis

In this section, we will analyze some of the mistakes made when applying the CoT-SC strategy in the selected models. Specifically, Table 4 refers to the errors made by the GPT model, and Table 5 refers to the errors found when using mistral.

For the errors of each GPT option, we will analyze the Spanish texts as they are the ones we understand the best. In the first example (id 1), we can see how the first neutral text option produces a grammatical text by removing the toxic words that represent the subject of the phrase "*Ah, porque se va y viene la*" (Ah, because the). In the second neutral option of that example we can see how it replaces the word "*subnormal*" (subnormal) with "*tonto*" (dumb) which is still a somewhat toxic word. Looking at example 2 (id 2) on the other hand, we can see that while options 1 and 2 start in a similar way "*y este*" (and this), option 3 changes its format more and eliminates the word "*y*" (and).

In the case of Mistral, we have selected two texts in which the errors detected are clearly visible. If we look at the table showing the errors made by the Mistral model (Table 5) we can see how in the example with id 3 the model does not generate the answer and in the example with id 4, it does not put the label "BEST PROPOSAL:", so we cannot identify the best-reasoned option.

5. Conclusion

This paper presents the participation of SINAI research group in the Multilingual Text Detoxification (TextDetox) shared task at CLEF 2024.

We conclude that the inclusion of self-consistency in the chain of thought prompting strategy helps the model to obtain higher performance by exploring different ways to detoxify texts, avoiding those that are ungrammatical and contain errors. However, the performance of our system varies greatly depending on the language to which it is applied since the knowledge that the models have in languages with few resources is minimal, and more effort should be invested to improve this. Finally, it is worth noting the importance of doing a manual review of the texts generated by a generative model such

Table 5
Errors finds in the generation of detoxified text using Mistral-7B-instruct-v0.2 model and the dev set of the TextDetox Dataset.

Id	Toxic Text	Response
3	Also, I have the fucking rights to advertise my youtube.com videos	Detected language: English Reasoning: The text contains several toxic expressions such as "fucking rights," "hater," and "blocking them." These expressions are offensive and disrespectful. To detoxify the text while maintaining the semantics and fluidity, we can rewrite it in a more polite and respectful way. Here are three alternatives: 1. I believe I have the right to promote my YouTube videos here. I understand if some people may not be interested, but I will continue to share them. 2. I'd like to share my YouTube videos with this community. I hope everyone finds value in them, but I understand if some prefer not to engage. 3. I appreciate the opportunity to share my YouTube content with all of you. I hope you find it enjoyable and informative, but I respect your decision if you choose not to view it.
4	wir brauchen das scheiß #Europa nicht ich bin #Deutscher und kein Europäer #btw2017 *(We don't need fucking #Europe I'm #German and not European #btw2017)*	To detect the language of the text, I identified the hashtags "#Europa" and "#Deutscher" which are in German. First alternative: We don't need Europe. I am German and not European. #BESTPROPOSAL Reasoning: The original text contains the offensive expression "scheiß" which translates to "shit" in English. To detoxify the text, we removed this expression and kept the rest of the text as it is, ensuring content fidelity. Second alternative: Europe is not necessary for us. I am German and not European. #BESTPROPOSAL Reasoning: We replaced the offensive expression "scheiß" with "not necessary" to maintain the semantics of the original text while ensuring neutrality. Third alternative: Europe is not essential for us. I am German and not European. #BESTPROPOSAL Reasoning: We used the expression "not essential" instead of "not necessary" to make the text even more neutral and avoid any ambiguity. The best proposal is the third alternative as it is the most neutral and avoids any offensive expressions while maintaining the content and fluidity of the original text.

as GPT (which is the one used in this experiment) because although the automatic metrics give you a reference of how well a system may work, it may not resemble the reality in terms of validity and quality of the texts.

In future work, we propose a deeper study to identify which toxic words are the most difficult for different LLMs to detect, which writing errors can invalidate these texts, and whether techniques such as True Zero-Shot learning can help these models better understand the grammatical rules of various languages. Additionally, we aim to explore the Tree of Thought strategy and other text detoxification methods to incorporate them into the LLMs' prompts, improving the instructions we provide to the models.

Acknowledgments

This work has been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00), and Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

References

[1] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Lang Resources and Evaluation 55 (2021) 477–523. doi:https://doi.org/10.1007/s10579-020-09502-8.

[2] F. M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña-López, M. T. Martín-Valdivia, A multi-task learning approach to hate speech detection leveraging sentiment analysis, IEEE Access 9 (2021) 112478–112489. doi:10.1109/ACCESS.2021.3103697.

[3] M. Subramanian, V. Easwaramoorthy Sathiskumar, G. Deepalakshmi, J. Cho, G. Manikandan, A survey on hate speech detection and sentiment analysis using machine learning and deep learning models, Alexandria Engineering Journal 80 (2023) 110–121. URL: https://www.sciencedirect.com/science/article/pii/S1110016823007238. doi:https://doi.org/10.1016/j.aej.2023.08.038.

[4] T. Bourgeade, Z. Li, F. Benamara, V. Moriceau, J. Su, A. Sun, Humans need context, what about machines? investigating conversational context in abusive language detection, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 8438–8452. URL: https://aclanthology.org/2024.lrec-main.740.

[5] N. Ejaz, F. Razi, S. Choudhury, Towards comprehensive cyberbullying detection: A dataset incorporating aggressive texts, repetition, peerness, and intent to harm, Computers in Human Behavior 153 (2024) 108123. URL: https://www.sciencedirect.com/science/article/pii/S0747563223004740. doi:https://doi.org/10.1016/j.chb.2023.108123.

[6] M. Anand, K. B. Sahay, M. A. Ahmed, D. Sultan, R. R. Chandan, B. Singh, Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques, Theoretical Computer Science 943 (2023) 203–218. URL: https://www.sciencedirect.com/science/article/pii/S0304397522003887. doi:https://doi.org/10.1016/j.tcs.2022.06.020.

[7] H. Bonaldi, Y.-L. Chung, G. Abercrombie, M. Guerini, Nlp for counterspeech against hate: A survey and how-to guide, 2024. arXiv:2403.20103.

[8] M. E. Vallecillo Rodríguez, M. V. Cantero Romero, I. Cabrera De Castro, A. Montejo Ráez, M. T. Martín Valdivia, CONAN-MT-SP: A Spanish corpus for counternarrative using GPT models, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and

Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 3677–3688. URL: https://aclanthology.org/2024.lrec-main.326.

[9] 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.

[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] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, 2023. URL: https://arxiv.org/abs/2203.11171. arXiv:2203.11171.

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, 2023. arXiv:2201.11903.

[13] M. T. I. Khondaker, M. Abdul-Mageed, L. V. S. Lakshmanan, Greenllama: A framework for detoxification with explanations, 2024. arXiv:2402.15951.

[14] 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. URL: http://dx.doi.org/10.1109/ICMLA58977.2023.00230. doi:10.1109/icmla58977.2023.00230.

[15] A. Parnami, M. Lee, Learning from few examples: A summary of approaches to few-shot learning, 2022. arXiv:2203.04291.

[16] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, Q. M. J. Wu, A review of generalized zero-shot learning methods, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2023) 4051–4070. doi:10.1109/TPAMI.2022.3191696.

[17] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. arXiv:2005.14165.

[18] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. arXiv:2310.06825.

[19] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

[20] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. URL: https://link.

springer.com/chapter/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6_20.
[21] 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. arXiv:1911.02116.
[22] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic bert sentence embedding, 2022. arXiv:2007.01852.
[23] M. Popović, chrF: character n-gram F-score for automatic MT evaluation, in: O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, P. Pecina (Eds.), Proceedings of the Tenth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 392–395. URL: https://aclanthology.org/W15-3049. doi:10.18653/v1/W15-3049.