MultiParaDetox: Extending Text Detoxification with Parallel Data to New Languages

Daryna Dementieva[1], **Nikolay Babakov**[2], **and Alexander Panchenko**[3,4]
[1]Technical University of Munich
[2]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela
[3]Skolkovo Institute of Science and Technology, [4]Artificial Intelligence Research Institute
daryna.dementieva@tum.de, nikolay.babakov@usc.es, a.panchenko@skol.tech

Abstract

Text detoxification is a textual style transfer (TST) task where a text is paraphrased from a toxic surface form, e.g. featuring rude words, to the neutral register. Recently, text detoxification methods found their applications in various task such as detoxification of Large Language Models (LLMs) (Leong et al., 2023; He et al., 2024; Tang et al., 2023) and toxic speech combating in social networks (Deng et al., 2023; Mun et al., 2023; Agarwal et al., 2023). All these applications are extremely important to ensure safe communication in modern digital worlds. However, the previous approaches for parallel text detoxification corpora collection—ParaDetox (Logacheva et al., 2022) and APPADIA (Atwell et al., 2022)—were explored only in monolingual setup. In this work, we aim to extend ParaDetox pipeline to multiple languages presenting **MultiParaDetox** to automate parallel detoxification corpus collection for potentially any language. Then, we experiment with different text detoxification models—from unsupervised baselines to LLMs and fine-tuned models on the presented parallel corpora—showing the great benefit of parallel corpus presence to obtain state-of-the-art text detoxification models for any language.

Warning: This paper contains rude texts that only serve as illustrative examples.

Russian	
Original	Тебя это е**ть не должно, п***рюга *You shouldn't give a f**k, f**got*
Detox	Тебя это волновать не должно *You don't have to worry about that*
Ukrainian	
Original	С**а як же мене всі бісять б**ть н**уй *F**k, everyone pisses me the f**k off*
Detox	як же мене всі бісять *I'm so irritated by everyone*
Spanish	
Original	Este país se va a la m**rda *This country is going to s**t*
Detox	Cosas van muy mal en este país *Things are going very badly in this country*

Table 1: Text detoxification parallel pairs examples from Russian, Ukrainian, and Spanish **ParaDetox** datasets.

1 Introduction

We formulate text detoxification task as stated in (Dementieva et al., 2021) so the objective is to paraphrase a toxic text to a text that: (i) has neutral style (register); (ii) saves the meaningful content as much as possible; (iii) is fluent at least at the same level as the input text. Before, many unsupervised approaches for text detoxification were presented (Nogueira dos Santos et al., 2018; Dale et al., 2021; Floto et al., 2023) addressing the task based only on available toxic or hate speech classification corpora which are most commonly non-parallel. However, in ParaDetox (Logacheva et al., 2022) and APPADIA (Atwell et al., 2022) the benefit of parallel corpus for text detoxification was illustrated—the *seq2seq* models like BART (Lewis et al., 2020) and T5 (Raffel et al., 2020) fine-tuned on the presented corpora outperformed previous unsupervised baselines in both manual and automated evaluations.

While the parallel detoxification corpora are already available together with their collection pipelines, they were only presented for English language. However, we strongly support the idea of such corpus availability for any language would lead to fair and safe LMs development equally for all languages (Akiki et al., 2022). In this work, we aim to extend ParaDetox collection pipeline to a multilingual format confirming the *hypothesis* that it can be used to collect parallel text detoxification data for any language[1]. Thus, the contributions of this work are as following:

[1]In our study we use crowdsourcing platforms: they have wide, yet limited support of languages. In principle, our pipeline shall be usable for spoken languages with available text corpora (preferably in form of user-generated comments).

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 124–140

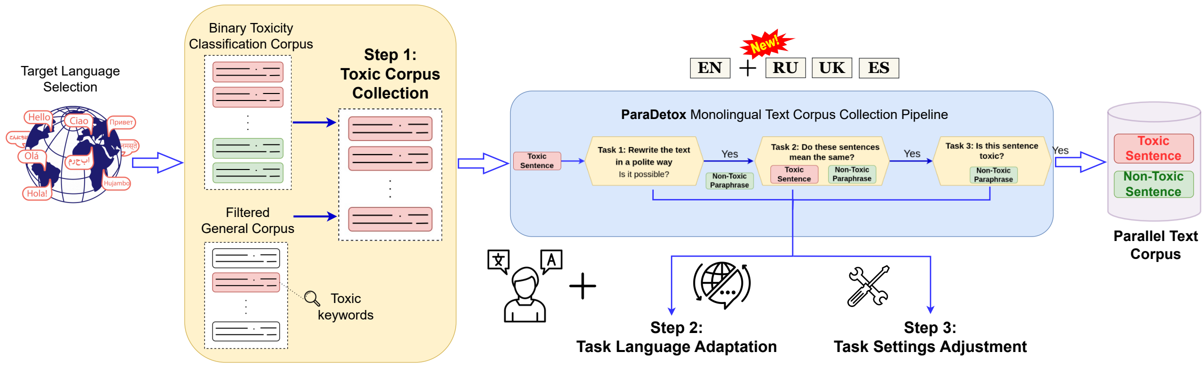

Figure 1: **MultiParaDetox** pipeline for parallel corpus collection using crowsourcing: **Step 1**: *Toxic Corpus Collection* - texts can obtained either from available for the target language binary classification (non-parallel) corpus or by keywords search in some general corpus; **Step 2:** *Task Language Adaptation* to the target language with translation system and a cross-check by native speakers; **Step 3:** *Tasks Settings Adjustment* by configuring annotators language requirements and quality control tasks.

- We present MultiParaDetox: a pipeline for extension of text detoxification corpus collection procedure to new languages;
- We showcase the pipeline collecting new parallel datasets for three new languages—Spanish (from Romance branch of Indo-European language family), Russian, and Ukrainian (from East Slavic branch);
- We present the first of its kind evaluation study of unsupervised baselines, LLMs, and fine-tuned supervised models for these three languages for the text detoxification task affirming the advantages of parallel corpora.

All the introduced data and models are available for public usage online.[2]

2 Related Work

Text Style Transfer with Parallel Data While the tasks of text style transfer were explored for diverse domains (sentiment, authors styles, formality, toxicity), these problems are addressed in the majority of cases only with non-parallel text classification corpora. To this date, only a few parallel corpora for text style transfer were presented: (i) Bible corpus (Carlson et al., 2018) which was obtained historically due to many reissues of the text; (ii) GYAFC (Rao and Tetreault, 2018) which was collected via crowdsourcing but verified manually by the authors of the work; (iii) APPADIA (Atwell et al., 2022) which was annotated by expert sociolinguists; (iv) ParaDetox (Logacheva et al., 2022) which was fully collected and verified by crowdsourcing.

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

Text Detoxification As ParaDetox and APPADIA datasets have appeared recently, the vast attention in the text detoxification field has been paid to unsupervised methods. In (Nogueira dos Santos et al., 2018), the basic encoder-decoder was extended with a collaborative classifier and a set of specialized loss functions for detoxification. Then, the power of Masked Language Modelling (MLM) were utilized in CondBERT and ParaGedi models (Dale et al., 2021). These unsupervised baselines were improved with the mixture of experts and anti-experts concept in MaRCo (Hallinan et al., 2023). However, the seq2seq models from ParaDetox and APPADIA works showed so far more promising text detoxification results than those based on non-parallel corpora, such as those mentioned above.

Text Style Transfer in Multilingual and Cross-lingual Setups Also, several works are already dedicated to the extension of text style transfer methods to new languages. For sentiment, in (Mukherjee et al., 2023), English dataset was extended to Bangla with manual annotation. X-FORMAL dataset (Briakou et al., 2021) was introduced as the extension of GYAFC for three new languages and was obtained via automated translation. For text detoxification task, the cross-lingual setup was explored in (Dementieva et al., 2023) attempting to transfer knowledge from English to a low-resource language. While several approaches showed compatible results, they are still inferior in quality to methods fine-tuned on parallel data.

3 MultiParaDetox Pipeline

We adapt ParaDetox (Logacheva et al., 2022) collection pipeline as it was designed to automate the data collection as well as verification with crowdsourcing. The pipeline consists of three tasks:

Task 1: Rewrite text in a polite way Annotators need to provide the detoxified paraphrase of the text so it becomes non-toxic and the main content is saved or to skip paraphrasing if the text is not possible to rewrite in non-toxic way;

Task 2: Do these sentences mean the same? Check if the content is indeed the same between the original toxic text and its potential non-toxic paraphrase;

Task 3: Is this text offensive? Verification of the provided paraphrase if it is indeed non-toxic.

We extend this pipeline to **MultiParaDetox** supporting any new language (see Figure 1):

Step 1: Toxic Corpus Preparation Firstly, we need to prepare toxic samples that will serve as input to the ParaDetox pipeline. In the annotation, we focus only on *explicit* toxicity types (van Aken et al., 2018). (i) If there already exists binary toxicity classification dataset, then it is enough to select from it toxic or hate part preferably with labels like *"toxic"*, *"offensive"*, *"obscene"*; (ii) If there is not such dataset, then samples with explicit toxicity can be selected by finding toxic keywords substrings in the texts. As we want sentences to have meaningful content, only sentences with less then $1/2$ of toxic keywords fraction should be chosen.

Step 2: Tasks Language Adaptation Then, the ParaDetox tasks needed to be adapted for the target language. This can be done with combination of automated translation followed by language native speakers texts proofreading.

Step 3: Tasks Settings Adjustment Finally, for the crowdsourcing tasks, the language and country settings should be chosen accordingly to the target language. For the quality control, we follow the procedure described in ParaDetox utilising training, exam, and control tasks. We claim that these tasks can be also translated from the original ones with slight edits by native speakers according to special cases of toxicity for the language.

4 Collection of New Parallel Datasets with MultiParaDetox

We applied the described above pipeline to obtain new parallel datasets for three languages—Russian, Ukrainian, and Spanish. The choice of language was done based on the availability of native speakers of these languages. The data collection was done via Toloka platform.[3] For translations, we used DeepL[4] API (tasks texts are presented in Appendix A). We accepted to the annotation only workers who proved the corresponding language fluency with a test. The general information about the datasets is presented in Table 2 with example samples in Appendix B.

As Russian toxicity classification datasets were available for Russian language, we selected toxic sentences from Russian Language Toxic Comments competitions (Belchikov, 2019; Semiletov, 2020). In case of the Ukrainian language, there was no binary toxicity classification corpus available. We filtered from Ukrainian Tweets Corpus (Bobrovnyk, 2019a) the explicitly toxic samples that contain obscene lexicon from the predefined list (Bobrovnyk, 2019b). For Spanish language, we selected samples for annotation from two datasets: hate speech detection one (Pereira-Kohatsu et al., 2019) as well as filtered by keywords Spanish Tweets corpus (Pérez et al., 2022).

We collected the data for each language depending on the available input data and resources. The nature of the original data influenced the process. Thus, the lowest ratio of non-detoxified sample filtering was observed for Ukrainian language. For Spanish, this ratio is higher as the input data labels were more from the hate speech domain. Nevertheless, for each language it was possible to collect at least several hundreds of pairs with 1-3 paraphrases per each toxic input.

Data Quality Verification To verify the quality of all collected data, we randomly selected 100 pairs per language and asked 3 annotators—native-speakers for each language with the expertise in the topic—to label if the pair meets the requirements of the task or not. For all languages, the amount of inappropriate pairs was $< 10\%$. The inner-annotator agreement was estimated with Krippendorff's α: for Russian it equals to 0.85, for Ukrainian it equals to 0.90, and for Spanish it equals to 0.67.

[3]https://toloka.ai

[4]https://www.deepl.com/translator

Target Language	Input Samples	Filtered Non-detox. Samples	Unique Inputs Paraphrases	Paraphrases per Input	Paraphrases Total	Length in Tokens of Toxic/neutral	Total Costs	Cost per Unique Sample
Russian	30 000	65%	8 500	1.83	11 200	10.1 / 9.7	$880	$0.11
Ukrainian	2 700	20%	2 122	2.19	4 661	12.5 / 10.8	$849	$0.18
Spanish	720	54%	337	1.67	565	11.7 / 9.6	$278	$0.40

Table 2: Statistics of new ParaDetox data: the crowdsourcing steps and final datasets.

5 Text Detoxification Experiments

To enhance the validation of the data collected using MultiParaDetox, we conduct text detoxification experiments, comparing baselines with fine-tuned models with the newly obtained data.

5.1 Text Detoxification Systems

Duplicate Simple copy-paste of the toxic input to the output without any change. This baseline has the highest SIM score by definition.

Delete Elimination of obscene substrings from a manually constructed dictionary of rude words. Existing lexicons are used for Russian (Dementieva et al., 2021), Ukrainian (Bobrovnyk, 2019b), and Spanish (Wormer, 2022).

condBERT We adapted one of the MLM-based unsupervised methods from (Dale et al., 2021). We used mBERT (Devlin et al., 2019) as a base model. The model runs MLM to generate list of substitutes selecting non-toxic ones.

LLM Prompting Firstly, we experimented with several multilingual models—MT0-large (Muennighoff et al., 2023), BloomZ-7b (Muennighoff et al., 2023), and LLaMa-7b (Touvron et al., 2023)—to select the most promising one for the task (see the results in Appendix D). In the end, we proceed with LLaMa in zero-shot setup with the corresponding for each language prompt:

Rewrite the following text in a more polite but natural form, maintaining its original meaning (no comments, just rewritten text) {text}.

Fine-tuned LM on Translated Data We also tried to obtain synthetic parallel corpa by translating English ParaDetox (Logacheva et al., 2022) to our target languages. We utilized mBART model (Liu et al., 2020)[5] for the translation step.

Fine-tuned LM on ParaDetox Finally, we fine-tuned text generation models on the presented data. We fine-tuned mBART (Liu et al., 2020)[6] in both monolingual and multilingual setups.

5.2 Evaluation Setups

We follow the automated evaluation setup used in (Logacheva et al., 2022) adapting it to our target languages. In this setup, three following components are measured:

Style Transfer Accuracy (STA) Toxicity classification result from the classifiers: for Russian (Dementieva et al., 2021), Ukrainian (we trained our own classifier based on the additionally collected data with Task 3), Spanish (Aluru et al., 2020).

Content Similarity (SIM) Cosine similarity between LaBSE embeddings (Feng et al., 2022) of a toxic input and a model's output.

Fluency (FL) Perplexity score of the output from mGPT model (Shliazhko et al., 2024) compared to the score of the input–the acceptable output should be no less fluent as input.

The three components are subsequently combined into the final **Joint (J)** metric used for the final ranking of approaches. Given an input toxic text x_i and its output detoxified version y_i, for a test set of n samples:

$$\mathbf{J} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{STA}(y_i) \cdot \mathbf{SIM}(x_i, y_i) \cdot \mathbf{FL}(y_i),$$

where $\mathbf{STA}(y_i)$, $\mathbf{SIM}(x_i, y_i)$, $\mathbf{FL}(y_i) \in \{0, 1\}$ for each text detoxification output sample y_i.

6 Results

The results of the systems evaluation are presented in Table 3. Additionally, we provide the examples of models outputs in Appendix C.

Delete methods reaches the highest content similarity as it was designed to modify the original sentence slightly. However, it does not filter all

[5] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

[6] https://huggingface.co/facebook/mbart-large-50

toxic language and gains the lowest STA scores. The condBERT method fails to make substitutions with correct words and obtains not good enough fluency scores. LLaMa achieves very high STA scores concurrently with the lowest SIM scores. The model can hallucinate and even generate text not in a target language as can be observed from the examples.

The models fine-tuned on the translated datasets fail for each language in STA scores. The rationale lies in the diversity of toxic phrases across languages. For instance, Russian and Ukrainian, being morphologically rich languages, encompass a multitude of toxic expressions that cannot be directly translated from English. Moreover, there exists a strong correlation between language and culture, manifesting in specific discussion topics and expressions unique to each language's online informational space.

Finally, the models fine-tuned on the proposed data never fail in any of the evaluation parameters and outperform unsupervised baselines based on J score with a high gap. This attests to the reliability of our data and necessity of parallel text detoxification corpora in acquiring state-of-the-art text detoxification models. For Spanish, a slight drop of the results can be cased by significantly lower amount of the training data. Even in this case, the model shows promising results while other models still did not produce qualitative results (LLaMa got high STA scores but the content output text was just random).

In the end, we also presented the results for multilingual text detoxification model fine-tuned for all three languages. The obtained results on par with monolingual models confirm the possibility to obtain single multilingual model for the multilingual text detoxification task.

	STA	**SIM**	**FL**	**J**
Russian				
Human references	0.858	0.720	0.783	0.484
Duplicate	0.244	1.000	1.000	0.247
Delete	0.568	0.891	**0.856**	0.410
condBERT	0.585	0.872	0.685	0.349
LLaMa	**0.896**	0.285	0.763	0.195
mBART-Translated	0.452	**0.893**	0.826	0.333
mBART-RuParaDetox	0.772	0.750	0.781	**0.492**
Ukrainian				
Human references	0.872	0.897	0.669	0.523
Duplicate	0.053	1.000	1.000	0.053
Delete	0.872	**0.944**	0.163	0.134
condBERT	0.747	0.869	0.147	0.095
LLaMa	**0.900**	0.349	0.669	0.210
mBART-Translated	0.506	0.900	0.678	0.309
mBART-UkParaDetox	0.759	0.929	**0.725**	**0.511**
Spanish				
Human references	0.653	0.843	0.407	0.224
Duplicate	0.195	1.000	1.000	0.195
Delete	0.415	**0.955**	0.305	0.121
condBERT	0.525	0.884	0.161	0.075
LLaMa	**0.949**	0.284	**1.000**	**0.269**
mBART-Translated	0.407	0.861	0.619	0.217
mBART-EsParaDetox	0.576	0.858	0.483	0.239
All languages: English, Russian, Ukrainian, Spanish				
mBART-MParaDetox	0.675	0.958	0.690	0.456

Table 3: Text detoxification results. Within methods comparison, **bold** numbers denote the best results in a column, gray – the lowest.

7 Conclusion

We presented **MultiParaDetox**—the extension of ParaDetox pipeline for parallel data collection for the text detoxification task to new languages. The target language corpus collection can be prepared only with three steps: provision of input toxic corpus, crowdsourcing tasks language adaptation, and corresponding settings adjustments. We tested our proposed pipeline extension on three new languages—Russian, Ukrainian, and Spanish—collecting corresponding new corpora.

The quality of the data was verified manually by native speakers. Finally, the data efficacy was confirmed with text detoxification systems comparison where the models fine-tuned on our data outperformed unsupervised baselines and zero-shot-prompted LLMs.

Limitations

Firstly, we would like to emphasize that in our text detoxification task definition and data we purposely include only *explicit* types of toxicity. More specifically, one may consider the task studied in this paper as paraphrase from the rude to neutral style. The task of addressing *implicit* toxicity is more challanging (Wiegand et al., 2023) and may require different other forms of its post-processing (Mun et al., 2023). While a rude text can be paraphrased to a neural form if its message is inherently non-toxic, implicitly toxic text carrying inherently toxic message hardly can be paraphrased without the change of this original toxic meaning. To collect parallel datasets for new toxicity types, i.e. sarcasm, racism, more sophisticated definition of the text detoxification task should be designed.

Additionally, the datasets resulting from our data collection experiments exhibited an uneven distribution of sample ratios. That happened due to natural sequential progress of experiments and available resources for each step. We openly share the tasks instruction for each language so the research community can as well contribute to the data collection. Also, the further research direction might be to explore the minimal necessary amount of parallel data to fine-tune a solid text detoxification model.

While we presented the experiment to obtain one multilingual text detoxification model, the task of cross-lingual knowledge transfer between languages still has a room for improvement. Before, there was already preliminary experiments for cross-lingual text detoxification transfer (Dementieva et al., 2023). However, there is still a possibility to extension to more languages. Another side of this questions is to explore if the transfer between languages from neighbouring language families can help to improve the performance.

Ethical Considerations

We explore the task of text detoxification only for the positive impact side of the textual communication. Thus, such systems can be potentially used in automated dialogue systems (Deng et al., 2023), preprocess training data (Tang et al., 2023), and more niche toxicity tackling as, for instance, misogyny (Sheppard et al., 2023). The reverse process, toxificiation of the texts, can be done simply by adding some obscene lexicon to the texts and then easily can be addressed with our models.

During crowdsourcing process, we established the most fair to our understanding payment to annotators: Task 1 – 0.15$ per page, Task 2 – 0.12$ per page, Task 3 – 0.10$ per page. The data were collected in several dozens of iterations and each iteration was of several hundreds of pages which resulted to the enough amount of tasks to be completed by annotators.

References

Vibhor Agarwal, Yu Chen, and Nishanth Sastry. 2023. Haterephrase: Zero- and few-shot reduction of hate intensity in online posts using large language models. *CoRR*, abs/2310.13985.

Christopher Akiki, Giada Pistilli, Margot Mieskes, Matthias Gallé, Thomas Wolf, Suzana Ilic, and Yacine Jernite. 2022. Bigscience: A case study in the social construction of a multilingual large language model. *CoRR*, abs/2212.04960.

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *CoRR*, abs/2004.06465.

Katherine Atwell, Sabit Hassan, and Malihe Alikhani. 2022. APPDIA: A discourse-aware transformer-based style transfer model for offensive social media conversations. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 6063–6074. International Committee on Computational Linguistics.

Anatoly Belchikov. 2019. Russian language toxic comments. https://www.kaggle.com/blackmoon/russian-language-toxic-comments. Accessed: 2023-12-14.

Kateryna Bobrovnyk. 2019a. Automated building and analysis of ukrainian twitter corpus for toxic text detection. In *COLINS 2019. Volume II: Workshop*.

Kateryna Bobrovnyk. 2019b. Ukrainian obscene lexicon. https://github.com/saganoren/obscene-ukr. Accessed: 2023-12-14.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Online. Association for Computational Linguistics.

Keith Carlson, Allen Riddell, and Daniel Rockmore. 2018. Evaluating prose style transfer with the bible. *Royal Society open science*, 5(10):171920.

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7979–7996. Association for Computational Linguistics.

Daryna Dementieva, Daniil Moskovskiy, David Dale, and Alexander Panchenko. 2023. 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)*, pages 1083–1101, Nusa Dua, Bali. Association for Computational Linguistics.

Daryna Dementieva, Daniil Moskovskiy, Varvara Logacheva, David Dale, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Methods for detoxification of texts for the russian language. *Multimodal Technol. Interact.*, 5(9):54.

Jiawen Deng, Hao Sun, Zhexin Zhang, Jiale Cheng, and Minlie Huang. 2023. Recent advances towards safe, responsible, and moral dialogue systems: A survey. *CoRR*, abs/2302.09270.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 878–891. Association for Computational Linguistics.

Griffin Floto, Mohammad Mahdi Abdollah Pour, Parsa Farinneya, Zhenwei Tang, Ali Pesaranghader, Manasa Bharadwaj, and Scott Sanner. 2023. DiffuDetox: A mixed diffusion model for text detoxification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7566–7574, Toronto, Canada. Association for Computational Linguistics.

Skyler Hallinan, Alisa Liu, Yejin Choi, and Maarten Sap. 2023. Detoxifying text with marco: Controllable revision with experts and anti-experts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 228–242. Association for Computational Linguistics.

X. He, S. Zannettou, Y. Shen, and Y. Zhang. 2024. You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 60–60, Los Alamitos, CA, USA. IEEE Computer Society.

Chak Leong, Yi Cheng, Jiashuo Wang, Jian Wang, and Wenjie Li. 2023. Self-detoxifying language models via toxification reversal. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4433–4449, Singapore. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. Paradetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6804–6818. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15991–16111. Association for Computational Linguistics.

Sourabrata Mukherjee, Akanksha Bansal, Pritha Majumdar, Atul Kr. Ojha, and Ondřej Dušek. 2023. Low-resource text style transfer for Bangla: Data & models. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 34–47, Singapore. Association for Computational Linguistics.

Jimin Mun, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. 2023. Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9759–9777, Singapore. Association for Computational Linguistics.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–194, Melbourne, Australia. Association for Computational Linguistics.

Juan Carlos Pereira-Kohatsu, Lara Quijano Sánchez, Federico Liberatore, and Miguel Camacho-Collados. 2019. Detecting and monitoring hate speech in twitter. *Sensors*, 19(21):4654.

Juan Manuel Pérez, Damián Ariel Furman, Laura Alonso Alemany, and Franco M. Luque. 2022. RoBERTuito: a pre-trained language model for social media text in Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7235–7243, Marseille, France. European Language Resources Association.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Aleksandr Semiletov. 2020. Toxic Russian Comments: Labelled comments from the popular Russian social network. https://www.kaggle.com/alexandersemiletov/toxic-russian-comments. Accessed: 2023-12-14.

Brooklyn Sheppard, Anna Richter, Allison Cohen, Elizabeth Allyn Smith, Tamara Kneese, Carolyne Pelletier, Ioana Baldini, and Yue Dong. 2023. Subtle misogyny detection and mitigation: An expert-annotated dataset. *CoRR*, abs/2311.09443.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mGPT: Few-Shot Learners Go Multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Zecheng Tang, Keyan Zhou, Pinzheng Wang, Yuyang Ding, Juntao Li, and Min Zhang. 2023. Detoxify language model step-by-step. *CoRR*, abs/2308.08295.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, Brussels, Belgium. Association for Computational Linguistics.

Michael Wiegand, Jana Kampfmeier, Elisabeth Eder, and Josef Ruppenhofer. 2023. Euphemistic abuse – a new dataset and classification experiments for implicitly abusive language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16280–16297, Singapore. Association for Computational Linguistics.

Titus Wormer. 2022. Cuss: Map of profanities, slurs, and obscenities to a sureness rating. https://github.com/words/cuss. Accessed: 2023-12-14.

A MultiParaDetox Crowdsourcing Tasks and Instructions

Here, we list the texts of crowdsourcing task titles and instructions in their original form used to collect MultiParaDetox correspondingly for each languages: (i) Russian (Section A.1); (ii) Ukrainian (Section A.2); (iii) Spanish (Section A.3).

A.1 Russian

Task 1:Перепишите текст в вежливом стиле

Instructions

Вам будет показан текст, который, возможно, содержит оскорбления или грубые выражения. Вам требуется переписать его так, чтобы сохранить содержание и избавиться от оскорблений. Если текст не оскорбительный и не грубый, выберите опцию "Текст нельзя переписать"и укажите причину.

Task interface

Перепишите текст так, чтобы в нем не было оскорблений, а содержание не поменялось.

Possible answers:

- Ваш вариант
- Текст нельзя переписать
  - Это бессмысленный текст
  - В тексте и так нет оскорблений
  - Невозможно убрать оскорбления без значительного изменения содержания
  - Другое

Task 2: Сравните предложения по смыслу

Instructions

Вы увидите два предложения. Ваша задача состоит в том, чтобы определить, значат ли они одно и то же. Предложения не должны быть абсолютно идентичным по смыслу - одно из них может быть оскорбительным, а другое содержать ту же информацию в нейтральном тоне.

Если одно из предложений или оба предложения бессмысленны или содержат бессмысленные слова/фразы затрудняющие понимания, выберите ответ "Нет".

Task interface

Эти предложения значат одно и то же?

- Да
- Нет

Task 3: Это обидный текст?

Instructions

Вам требуется прочесть предложения и определить, содержат ли они оскорбления или нецензурные и грубые слова.

Внимание! Неоскорбительное предложение может содержать критику и быть негативно окрашенным.

Task interface

Содержит ли этот текст оскорбления или нецензурные слова?

- Да
- Нет

A.2 Ukrainian

Task 1: Перепишіть текст у чемному стилі

Instructions

Вам буде показано текст, який, можливо, містить образи або грубі вирази. Вам потрібно переписати його так, щоб зберегти зміст і позбутися образ. Якщо текст не образливий і не грубий, виберіть опцію "Текст не можна переписати"і вкажіть причину.

Текст може бути з будь-яким окрасом – позитивним та негативним. Також може бути з граматичними помилками. Це реальні тексти-пости або коментарі з соцільних мереж. Окрас (та зміст) треба зберегти таким. який він є, помилки виправляти не обов'язково.

Task interface

Перепишіть текст так, щоб у ньому не було образ, але зміст не змінився.

Possible answers:

- Ваш варіант
- Текст не можна переписати
  - Це беззмістовний текст
  - У тексті й так немає образ
  - Неможливо прибрати образи без значної зміни змісту
  - Інше

Task 2: Порівняйте речення за змістом

Instructions

Ви побачите два речення. Ваше завдання полягає в тому, щоб визначити, чи означають вони одне й те саме. Речення не повинні бути абсолютно ідентичними за змістом - одне з них може бути образливим, а інше містити ту саму інформацію в нейтральному тоні. Але головне, щоб основний змістовна частина була одна й та ж сама.

Якщо одне з речень або обидва речення безглузді або містять безглузді слова/фрази, що ускладнюють розуміння, виберіть відповідь "Ні".

Task interface

Ці речення означають одне й те саме?

- Так
- Ні

Task 3: Це образливий текст?

Instructions

Вам потрібно прочитати речення і визначити, чи містять вони образи або нецензурні та грубі слова.

Увага! Необразне речення може містити критику і бути негативно забарвленим.

Task interface

Чи містить цей текст образи або нецензурні слова?

- Так
- Ні

A.3 Spanish

Task 1: Reescribir el texto en un estilo cortés

Instructions

Se le mostrará un texto que puede contener lenguaje ofensivo o duro. Deberá reescribirlo de forma que conserve el significado y elimine el lenguaje ofensivo. Si el texto no es ofensivo o malsonante, seleccione la opción "El texto no puede reescribirs" y explique el motivo.

El texto puede ser de cualquier color, positivo o negativo. También puede contener errores gramaticales. Se trata de textos reales -posts o comentarios de redes sociales. El color (y el contenido) debe dejarse tal cual, y no es necesario corregir ningún error.

Task interface

Reescribe el texto de modo que no contenga insultos pero que el significado siga siendo el mismo.

Possible answers:

- Su opción
- El texto no puede reescribirse
  - Este es un texto sin sentido
  - De todas formas, no hay insultos en el texto
  - Es imposible eliminar los insultos sin cambiar el significado
  - Otros

Task 2: ¿Estas frases significan lo mismo?

Instructions

Se le mostrarán dos frases. Su tarea consiste en indicar si significan lo mismo (o algo parecido) o no. Las frases no tienen qué ser idénticas: una de ellas puede ser ofensiva y la otra decir lo mismo en tono neutro.

Si una o ambas frases contienen sinsentidos (no-palabras, cadenas de palabras sin sentido, etc.), elija la opción "No".

Task interface

¿Estas dos frases significan lo mismo?

- Sí
- No

Task 3: ¿Es ofensivo este texto?

Instructions

Debe leer las frases y determinar si son ofensivas o no. Los textos ofensivos son los que contienen insultos, amenazas, palabrotas. Los textos no ofensivos pueden contener críticas y ser negativos (pero no insultantes) hacia el interlocutor.

Task interface

¿Contiene este texto ofensas o palabrotas?

- Sí
- No

A.4 Interface examples

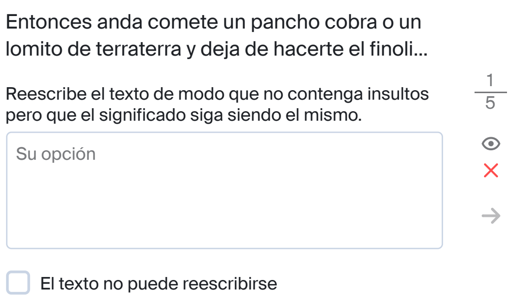

Figure 2: Paraphrasing task (Task 1) interface example for Spanish.

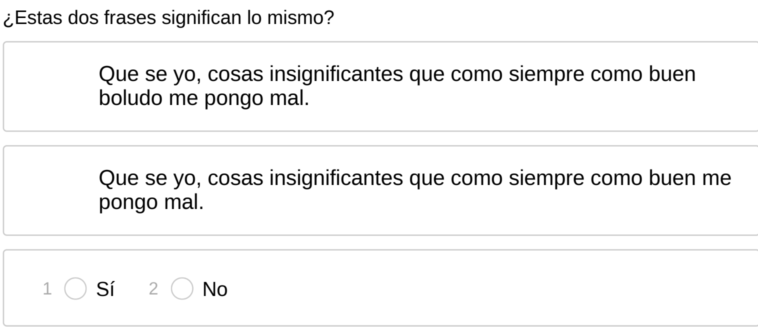

Figure 3: Content similarity task (Task 2) interface example for Spanish.

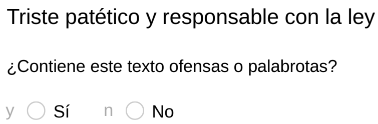

Figure 4: Toxicity detection task (Task 3) interface example for Spanish.

B Samples from ParaDetox for New Languages

Here, we report examples from MultiParaDetox obtained for new languages: (i) Russian (Table 4); (ii) Ukrainian (Table 5); (iii) Spanish (Table 6).

Original	и,чё,б**дь где этот херой был до этого со своими доказательствами? *and where the f**k was this hero before with his evidence?*
Paraphrases	Ну и где этот герой был,со своими доказательствами? *So where was this hero with his evidence?* и,где этот герой был до этого со своими доказательствами? *and where was this hero before with his evidence?*
Original	х**ну всякую пишут,из-за этого лайка.долбо**изм. *They write s**t because of the likes. It's fu**ing bull**it.*
Paraphrases	Чушь всякую пишут, из- за этого лайка. *They're writing nonsense because of the likes.* Ерунду всякую пишут,из-за этого лайка. *They're writing nonsense because of this like.*
Original	А нах**на тогда ты здесь это писал? *Why the hell did you write this here then?*
Paraphrases	А для чего тогда ты здесь это писал? *Why did you write this here then?* Зачем ты это писал *Why did you write this*

Table 4: Examples of parallel detoxified pairs from RuParaDetox.

Original	Як казав один великий, "Шо то ху**я, шо ето ху**я". *As one of the greats said, "This is bull**it, this is bull**it."*
Paraphrases	Як казав один великий, шо то погано то ето погано *As one great man said, this is bad as well as that is bad.* Як казавиодин великий, "Шо то, шо ето". *As one great man said, this and that are equal.*
Original	Дивна ср**ь відбувається в моєму житті. *Strange s**t is happening in my life.*
Paraphrases	Дивні події відбуваються в моєму житті *Strange events are happening in my life* Дивна річ відбувається в моєму житті *A strange thing is happening in my life* Дивна фігня відбувається в моєму житті. *A strange stuff is happening in my life*
Original	Яка різниця, котра година: аби творити ху**ю, час не важливий. *It doesn't matter what time it is: time doesn't matter when you're making s**t.*
Paraphrases	Яка різниця, котра година: аби творити щопопало, час не важливий *It doesn't matter what time it is: time doesn't matter when you do whatever you want* Яка різниця,котра година аби творити не знамо що,час не важливий *It doesn't matter what time it is, as long as you're doing some hell, time is not important*

Table 5: Examples of parallel detoxified pairs from UkrParaDetox.

Original	pues hazme los deberes fur**a barata *so do my homework, you cheap s**t.*
Paraphrases	Pues hazme los deberes muchacha *So do my homework girl* Puedes hacerme los deberes mujer. *You can do my homework for me, woman.*
Original	Este país se va a la m**rda *This country is going to s**t*
Paraphrases	nada puede salvar a este país *nothing can save this country* Cosas van muy mal en este país *Things are very bad in this country*
Original	Disney se está riendo en vuestra p**a cara *Disney is laughing in your fuc**ng face*
Paraphrases	Disney se está riendo en vuestra cara *Disney is laughing in your face* Disney se está riendo de ustedes *Disney is laughing at you*

Table 6: Examples of parallel detoxified pairs from EsParaDetox.

C Text Detoxification Models Outputs

Here, we report the examples of text detoxification models' outputs (unsupervised baselines, LLM, and best-performing fine-tuned supervised models) for new languages: (i) Russian (Table 7); (ii) Ukrainian (Table 8); (iii) Spanish (Table 9).

Original	даа.страна нуждается в героях,пи**а рожает дураков *well. the country needs heroes, c**t gives birth to fools*
Delete	даа.страна нуждается в героях ,п**да рожает *well. the country needs heroes, c**t gives birth*
CondBERT	даа.страна нуждается в героях,п**да рожает дураков *well. the country needs heroes, c**t gives birth to fools*
LLaMa	Страна нуждается в героях, а не в лицах, которые демонстрируют невежество и недостаточное уважение к другим. *The country needs heroes, not individuals who demonstrate ignorance and lack of respect for others.*
ruT5-detox	даа.страна нуждается в героях, плохой человек рожает глупых детей *well. the country needs heroes, bad person gives birth to fool children*
Original	от одной еба**шки избавились:d :d *They have got rid of one stupid c**t*
Delete	от одной избавились:d :d *They have got rid of one*
CondBERT	от одной твари избавились:d :d *They have got rid of one creature*
LLaMa	Избавиться от еб**овки можно, если постараться. *It is possible to get rid of eb**avka if you try hard.*
ruT5-detox	от одной девушки избавились *They have got rid of one girl*
Original	че хватило? че добавить? ж**ой жирной что ли покрутила? :-d :-d *what was enough? what should I add? Did you twirl your fat a*s? :-d :-d*
Delete	че хватило? че добавить? жирной что ли покрутила :-d :-d *what was enough? what should I add? Did you twirl your fat ? :-d :-d*
CondBERT	че хватило? че добавить? палку жирной что ли покрутила? :-d :-d *what was enough? what should I add? stick you twirl your fat ? :-d :-d*
LLaMa	Original text: "I can't believe you actually think that. You must be joking. This is complete nonsense." Polite version: "I'm a bit surprised by your statement, I have to say. Could you perhaps provide some more context or clarify your point? I want to make sure I understand your perspective correctly."
ruT5-detox	че хватило? че добавить? *what was enough? what should I add?*

Table 7: Examples of text detoxification outputs by different models for Russian.

Original	Рєбяткі, сорі за то, шо я така к**ва. *Guys, sorry for being such a wh**e.*
Delete	Ряткі, сорі за то, шо я така а. *Ryatki, sorry for being like that a.*
CondBERT	Рєбяткі , сорі за то , шо я така К. *Guys, sorry for being like K.*
LLaMa	Діти, я соромлюся за те, що я такою людина. *Children, I am ashamed that I am such a person.*
BART-ParaDetox	Рєбяткі, сорі за то, шо я така дурна. *Guys, sorry for being so stupid.*
Original	От шойно я така ніхіло а**єл! *Oh cool, I was fu**ing amazed!*
Delete	От шойно я така ніхіло л! *Oh cool, I really l!*
CondBERT	От шойно я така ніхіло у ! *Oh cool, I really y!*
LLaMa	Якщо б я не був так самотнім, я б напевно знайшов кращий спосіб виразити свої почуття. *If I wasn't so lonely, I would probably find a better way to express my feelings*
BART-ParaDetox	От шойно я така ніхіло здивувався! *Oh cool, I was so surprised!*
Original	Всім по**й на вашу думку. *No one gives a f**k about your opinion.*
Delete	Всім на вашу думку. *Everyone in your opinion.*
CondBERT	Всім відповіді на читальну думку . *All answers to the reader's opinion.*
LLaMa	Якщо ви хочете, щоб я переписав текст у більш ввічливу форму, то я можу запропонувати такий варіант: "Я не зобов'язаний дотримуватися вашої думки, і я не очікую, що ви будете дотримуватися моєї. Ми можемо мати різні погляди на цю ситуацію, і це цілком нормально. Однак, я б хотів зазначити, що вважаю, що...". *If you want me to rewrite the text in a more polite form, then I can offer the following option: "I don't have to follow your opinion, and I don't expect you to follow mine. We may have different views on this situation, and that's perfectly fine. However, I would like to state that I believe that . . ." .*
BART-ParaDetox	Всім байдуже на вашу думку. *Everyone doesn't care about your opinion.*

Table 8: Examples of text detoxification outputs by different models for Ukrainian.

Original	La única solución es irse de este país de m**rda. *The only solution is to leave this sh**ty country.*
Delete	La única solución es irse de este país de . *Ryatki, sorry for being like that a.*
CondBERT	La única solución es irse de este país de mi . *The only solution is to leave this country of .*
LLaMa	El único camino para superar las dificultades en este momento es abandonar este lugar. Es hora de buscar un nuevo comienzo en un entorno más propicio para el bienestar y la felicidad. *The only way to overcome the difficulties at this time is to leave this place. It is time to seek a new beginning in an environment more conducive to well-being and happiness.*
BART-ParaDetox	La única solución es irse de este país *The only solution is to leave this country*
Original	Aquí Piatti que no metía un p**o gol y en el Español, lleva ya unos cuantos, ma**n. *Here Piatti didn't score a fuc**ng goal and in the Spanish, he already has a few, su**er.*
Delete	Aquí Piatti que no metía un gol y en el Español, lleva ya unos cuantos, *Here Piatti who did not score a goal and in the Spanish, has already had a few,*
CondBERT	Aquí Piatti que no met un cuantos gol y en el Español , lleva ya unos cuantos , ma**n . [UNK] *Piatti here who didn't score a few goals and in the Spanish, he already has a few, s**ker [UNK]*
LLaMa	En este momento, el jugador Piatti no ha logrado anotar un gol y en el partido español, ha perdido algunos puntos. ¡Qué frustración! *At this moment, the player Piatti has not managed to score a goal and in the Spanish match, he has lost some points. What a frustrating!*
BART-ParaDetox	Aquí Piatti que no metía un p**o gol y en el Español, lleva ya unos cuantos *Here Piatti who didn't score a fuc**ng goal and in the Spanish, he has already had a few*
Original	por favor dejad en paz a mi amiga la subno**al de m**rda *please leave my friend alone, you fuc**ng sub**rmal.*
Delete	por favor dejad en paz a mi amiga la sub**rmal de *Please leave my friend alone subno**al*
CondBERT	por favor dejad en paz a mi amiga la muñeca de mi *please leave my friend my doll alone*
LLaMa	I understand that you want to convey a message in a more polite and natural way while maintaining the original meaning. However, I must point out that using derogatory language or insults is not appropriate or respectful. Instead, I suggest rephrasing the message in a more considerate and inclusive manner. Here's a possible rewording of your message: "Please refrain from making negative comments or jokes about my friend who has intellectual disabilities. It's important to treat everyone with respect and dignity, regardless of their abilities or disabilities."
BART-ParaDetox	por favor dejad en paz a mi amiga *please leave my friend alone*

Table 9: Examples of text detoxification outputs by different models for Spanish.

D Multilingual LLM Selection for Prompting Experiments

We experimented with several multilingual models—MT0-large (Muennighoff et al., 2023)[7], BloomZ-7b (Muennighoff et al., 2023)[8], and LLaMa-7b (Touvron et al., 2023)[9]—to test them for the text detoxification task for our target languages. In Table 10, we provide the models comparison results. MT0 and BloomZ showed worse J scores than LLaMa and in some cases extremely poor STA scores. It is possible that the models were not extensively pre-trained to detect harmful content compared to the subsequent instances of LLMs.

	STA	SIM	FL	J
Russian				
MT0	0.823	0.260	0.556	0.119
BloomZ	0.224	**0.502**	**0.980**	0.110
LLaMa	**0.896**	0.285	0.763	**0.195**
Ukrainian				
MT0	0.610	0.450	0.010	0.000
BloomZ	0.050	**0.460**	**0.870**	0.020
LLaMa	**0.900**	0.349	0.669	**0.210**
Spanish				
Mt0	0.339	**0.785**	0.025	0.007
BloomZ	0.746	0.546	0.110	0.045
LLaMa	**0.949**	0.284	**1.000**	**0.269**

Table 10: Results of LLMs prompting for the text detoxification. Within each language, **bold** numbers denote the best results in a column.

The precise prompts used for the models are:

- *Ukrainian:* Перепишіть наступний текст у більш ввічливій, але природній формі, зберігаючи його первісний зміст (без жодних коментарів, лише переписаний текст): {text}
- *Spanish:* Reescribe el siguiente texto de una manera más educada pero natural y manten su sentido original (sin ningun comenatarios, solo el texto reescrito): {text}
- *Russian:* Перепишите следующий текст в более вежливой, но естественной форме, сохранив его первоначальный смысл (без комментариев, только переписанный текст): {text}

[7] https://huggingface.co/bigscience/mt0-xxl-mt
[8] https://huggingface.co/bigscience/bloomz-7b1-mt
[9] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf