Leveraging Comparable Toxicity Lexicons in Prompt Instructions for Multilingual Text Detoxification

Yassir El Attar[1], **Esra Dönmez**[1,2], **Nina Ohlendorf**[1], **Agnieszka Falenska**[1,2]
[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]Interchange Forum for Reflecting on Intelligent Systems, University of Stuttgart, Germany
{yassir.el-attar, esra.doenmez, nina.ohlendorf, agnieszka.falenska}@ims.uni-stuttgart.de

Abstract

To mitigate the prevalence of toxic language on digital social media, various NLP approaches have been proposed for automatic text detoxification. However, the potential of toxic expressions lexicons as a comparable cross-lingual resource to guide this process remains largely unexplored. In this work, we investigate how such resources can be effectively used to inform multilingual language models about what should and should not be considered *toxic*. We evaluate four models under two settings—zero-shot prompting and fine-tuning—to assess the impact of incorporating toxic expressions in prompt instruction, including in cross-lingual transfer scenarios. Our results show that both zero-shot prompting and fine-tuning approaches benefit considerably from adding toxic expressions in prompt instructions during training and/or inference. Our findings demonstrate that comparable, lightweight, language-specific toxic expressions lexicons constitute an effective mechanism for injecting explicit information about lexical toxicity into multilingual language models.

Keywords: text detoxification, multilinguality, cross-lingual transfer, comparable corpora, low-resource languages

1. Introduction

Disclaimer: certain figures and examples include potentially offensive content.

Toxic language, following the criteria by Dementieva et al. (2024b), is defined as text containing vulgar or profane content, regardless of whether it directly targets or insults individuals or groups. For instance, a message such as *"I f*cking love this movie!!"* is toxic due to its use of profane language, yet it carries no hateful or offensive intent.

Toxic content is highly prevalent on the internet, especially on social media and in online forums (Vasist et al., 2023; Radfar et al., 2020). It is known to be harmful to people's mental well-being (Waldron, 2012) and specifically affects minority groups (Thomas et al., 2021) and children (Breckheimer, 2001). These potential harms motivate the *text detoxification* task, an automatic mitigation approach defined as a form of text style transfer (Dale et al., 2021) in which the vulgar style of a message is neutralized while its meaning is kept intact. For instance, toxic *"I don't give a sh*t about your opinion!"* could be detoxified into *"I don't care about your opinion!"*: the style changes, but the message stays the same.

The feasibility of the detoxification task increased with the introduction of Large Language Models (LLMs) (Logacheva et al., 2022). However, despite this advancement, detoxification remains a challenging task. This challenge is evidenced by the fact that language models are often pretrained on filtered data in which toxic content has been removed (Mendu et al., 2025). While this filtering is intended to reduce harmful outputs, it may also impair the models' ability to recognize profane or abusive expressions. The most common strategy for addressing this problem is to fine-tune models on large collections of toxic–detoxified text pairs. However, this approach requires substantial amounts of annotated data, which are costly to create. Consequently, such resources are often unavailable for low-resource languages, and multilingual coverage frequently depends on machine-translated corpora (Rykov et al., 2024), potentially introducing additional noise and bias.

A recent line of work has explored a complementary approach: using **comparable, language-specific lexicons of toxic expressions** (e.g., swear words) to inform LM-based detoxification. These corpora consist of lexicons constructed independently across languages around the same conceptual domain, rather than obtained through direct translation (Dementieva et al., 2024b). Importantly, they are considerably less resource-intensive to construct than parallel toxic–detoxified datasets. However, there is currently no consensus on how such corpora should be integrated into detoxification pipelines. Existing approaches include detecting toxic expressions and removing them directly from text (Dementieva et al., 2024b), risking loss of meaning, or masking them for the model to replace (Nuthakki et al., 2025), which can result in unnatural outputs. In a more similar work to ours, Lai-Lopez et al. (2025) proposed tagging toxic expressions

Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026, pages 108–118
11 May 2026.

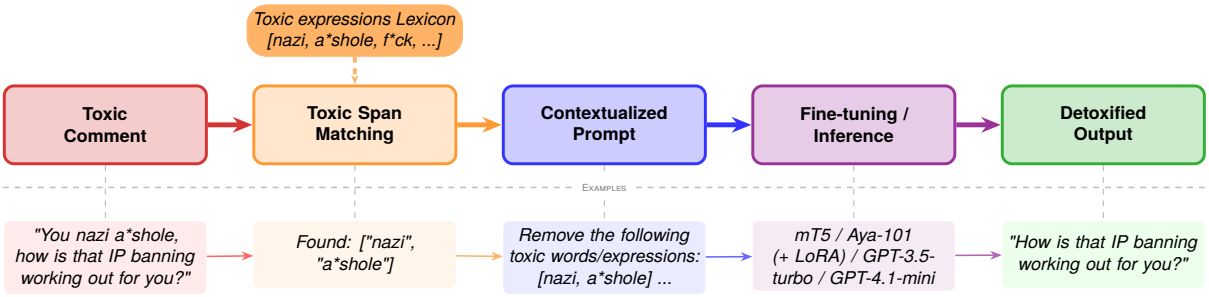

Figure 1: Overview of the toxic-expressions-in-instruction strategy. The upper row illustrates the processing pipeline stages, while the lower row provides a concrete example for each stage. The toxic expressions lexicon feeds into the matching/lookup stage to identify toxic expressions before using them to create our toxic-contextualized prompt instruction.

in inputs via markup (`<toxic>...</toxic>`). Yet, encoding lexical toxicity through such input annotations rather than through explicit instructions leaves open the question of whether Language Models (LMs) can be more effectively guided by directly specifying which expressions are toxic.

In this work, we ask *how to effectively inform language models* about what should and should not be considered *toxic*. We hypothesize that providing toxic expressions from comparable lexicons alongside the input sentence, together with explicit instructions to remove or replace them while preserving meaning, can lead to more controlled and semantically faithful detoxification. This motivates the following research questions:

RQ1: Does providing toxic expressions in model prompt instructions improve detoxification performance across model adaptation settings?

RQ2: To what extent do these improvements generalize in a cross-lingual setup, particularly for low-resource languages not seen during training?

To answer these questions, we evaluate the performance of four multilingual language models with and without toxic expressions in prompt instructions. Figure 1 demonstrates the overall pipeline with an example at each stage. Evaluation covers two model-adaptation settings: zero-shot prompting and fine-tuning. All experiments are evaluated on 15 languages from diverse typological families. In the case of fine-tuned models, evaluation additionally assesses cross-lingual generalization: the models are trained on data from 9 languages and then evaluated both on those languages and on 6 additional unseen languages. We find that across all strategies, model sizes, and language settings, providing toxic expressions in instructions consistently improves detoxification performance (Section 4). Moreover, the proposed method achieves remarkable results, demonstrating that this simple instruction-based knowledge injection is competitive with more data-intensive or architecture-specific approaches. Our results highlight the value of investing in the creation and curation of multilingual toxic expressions lexicons as a comparable resource. These resources can serve as a general mechanism for injecting domain-specific conceptual knowledge into language models, with potential benefits extending well beyond the detoxification task. Our experimental code and scripts are publicly available on GitHub[1].

2. Related Work

Toxic language is commonly conflated in the literature with other related concepts (Fortuna et al., 2020). However, following the definition by Dementieva et al. (2024b), it differs from broader notions such as hate speech, which targets individuals or groups based on characteristics such as race, gender, or religion (Davidson et al., 2017; Basile et al., 2019), or offensive language, which encompasses a wider range of socially unacceptable expressions that may not involve profanity (Fortuna et al., 2020). As Fortuna et al. (2020) highlight, these categories refer to distinct phenomena and require different methodological approaches for their detection and/or removal.

In order to prevent digital violence (Shi et al., 2020) and maintain constructive communication, AI models have been developed to detect (D'Sa et al., 2020; Zampieri et al., 2020), delete (Dementieva et al., 2024b) or block (Cobbe, 2021) toxic language. This detoxification process is a text style transfer (TST) task (Dale et al., 2021): The source style to be changed is the harmful toxic language, which is automatically transformed to the target style, the non-toxic, neutral language counterpart (Mukherjee et al., 2023a). Beyond the style change, the primary objective is to generate text that is fluent and preserves the original text's meaning as much as possible (Dementieva et al., 2021). As a supervised sequence-to-sequence task, this can be per-

[1] https://github.com/YassirELATTAR/multilingual-text-detoxification

	Language	Train	Test	# Toxic expressions
Seen languages	English	19,744 + 400	600	3,390
	Russian	12,206 + 400	600	141,000
	Ukrainian	400	600	7,360
	German	400	600	247
	Arabic	400	600	430
	Spanish	400	600	1,200
	Hindi	400	600	133
	Chinese	400	600	3,840
	Amharic	400	600	245
	Total	**35,550**	**5,400**	**157,845**
Unseen	French	—	600	1,290
	Italian	—	600	815
	Japanese	—	600	328
	Hinglish	—	600	209
	Tatar	—	600	15,600
	Hebrew	—	600	731
	Total	**—**	**3,600**	**18,973**

Table 1: Dataset statistics and toxic expressions lexicon size per language. More details on the sources and the data collection process can be found in Dementieva et al. (2025, 2024a); Logacheva et al. (2022); Dementieva et al. (2024b).

formed using encoder-decoder models trained on parallel data (Logacheva et al., 2022). Although unsupervised approaches exist (Nogueira dos Santos et al., 2018; Floto et al., 2023), supervised methods leveraging parallel corpora have proven particularly effective (Logacheva et al., 2022; Atwell et al., 2022). Subsequent work has focused on fine-tuning sequence-to-sequence models (Zhang et al., 2024), with approaches ranging from mT0 fine-tuning (Dementieva et al., 2024a) and GPT-4 few-shot prompting (Dementieva et al., 2025) to LoRA-based fine-tuning of Gemma-3 (12B) as the current state-of-the-art (Dang and D'Elia, 2025). However, to our knowledge, no prior work has systematically investigated toxic expressions lexicons as comparable corpora across both fine-tuning and prompting paradigms, nor evaluated their cross-lingual transferability to unseen languages, which is the gap the present work addresses.

3. Experimental Setup

We first describe the data resources, including datasets of toxic inputs paired with detoxified target rewrites and comparable multilingual toxic lexicons. Next, we detail experiments covering text detoxification settings, the toxic expressions matching and instruction construction procedures, and the evaluation metrics.

3.1. Data Resources

We make use of two resources: (1) **datasets of toxic comments paired with non-toxic (neutral) rewrites**, with training and test splits in 9 languages (see Table 2 for a few examples and Table 1 for dataset statistics) and additional test sets in six languages, and (2) **a comparable multilingual toxicity lexicon**[2] covering all 15 languages.

Datasets The multilingual dataset (Dementieva et al., 2025) includes training and test splits, with languages grouped into **seen** and **unseen** (as shown in Table 1). It provides 400 training instances per seen language[3] (English, Russian, Ukrainian, German, Arabic, Spanish, Hindi, Chinese, and Amharic) and 600 test instances per language for all seen and unseen languages: French, Italian, Japanese, Hinglish (in Latin alphabet), Tatar, and Hebrew[4].

Additionally, for training, we include English data from Logacheva et al. (2022)[5] ($19,744$ instances), and Russian from Dementieva et al. (2024a)[6] ($12,206$ instances) as shown in details in Table 1[7]. All instances are pairs of toxic comments and their detoxified (neutral) rewrites. After adding high-resource data (English and Russian in our case), the resulting training set becomes imbalanced, reflecting real-world differences in data availability across languages. For a detailed description of the datasets and their collection process, see Dementieva et al. (2025, 2024a,b) and Logacheva et al. (2022). In summary, the input and target pairs were obtained using a collection pipeline (Logacheva et al., 2022) in which human annotators were instructed to manually rewrite each toxic comment into a non-toxic paraphrase, verifying that the target rewrite is (1) non-toxic, (2) fluent, but may contain some minor mistakes depending on the input, and (3) semantically faithful to the original content. Any user names and links were anonymized.

Comparable multilingual toxicity lexicon (Dementieva et al., 2024b) The lexicon is a collection of toxic expressions across 15 languages ($176,818$ instances). It was compiled from multiple sources, existing community-maintained toxic expressions lists for most languages, and manually curated lists for Amharic and Arabic where no such resources existed. Additionally, the Tatar lexicon was created by merging an existing list with Russian toxic expressions translated into Tatar. The lexicons vary considerably in size across languages, rang-

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

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

[4] https://huggingface.co/datasets/textdetox/multilingual_paradetox_test

[5] https://huggingface.co/datasets/s-nlp/paradetox

[6] https://huggingface.co/datasets/s-nlp/ru_paradetox

[7] Both of these additional datasets come from the same underlying resource.

Lang.	Toxic comment	Neutral comment
English	lol i'm just f*ckin with ya!	lol i'm just playing with ya!
Spanish	Este país se va a la m*erda *(this country is going to sh*t)*	nada puede salvar a este país *(nothing can save this country)*
German	Weit und breit kein N*ger. *(Not a single n*gro in sight.)*	Weit und breit kein Schwarzer. *(Not a single black person in sight.)*

Table 2: Example pairs of toxic comments and non-toxic rewrites in the dataset from three different languages. The Spanish and German comments are accompanied by their translation.

Instruction-only	Toxic-expressions-in-instruction
Detoxify the sentence: "lol i'm just f*ckin with ya!".	Remove the following toxic words/expressions [f*ckin] from the sentence: "lol i'm just f*ckin with ya!", but keep the meaning and style similar to the original sentence.

Table 3: Prompts used in both zero-shot and fine-tuning settings. Instruction-only (w/o) simply instructs the model to detoxify the sentence, while toxic-expressions-in-instruction (w/) provides toxic expressions (words/expressions) as context in the input instruction.

ing from 133 entries for Hindi to $141,000$ for Russian (Table 1), reflecting the differences in resource availability rather than actual differences in toxic language use.

3.2. Experiments

We design and experiment with two settings (aka. model adaptations): **zero-shot prompting** and **fine-tuning**. In both settings, we employ two instruction strategies to test the effects of toxic expressions in prompt instructions on the model performance: **(a) instruction-only** (left column in Table 3; simply instructs models to detoxify the input based on its knowledge and the input data alone) and **(b) toxic-expressions-in-instruction** (right column in Table 3; explicitly provides the identified toxic expressions to guide the model in locating and handling them). To ensure a realistic scenario, we provide the instructions in the language of the input text (obtained via machine translation[8]).

Toxic expressions in instructions We inject toxic expressions into input instructions by leveraging the multilingual comparable lexicon of toxic expressions across all 15 languages. This step presents notable challenges for low-resource languages such as Tatar and Amharic, as well as for languages with distinct orthographic properties, such as Chinese, Arabic, and Hebrew, where simple string matching is insufficient due to the absence of word boundaries or complex morphology. To address this, we implement a language-aware matching tool that assigns a dedicated lookup function to each language. For whitespace-separated languages like English, Spanish, and German, we rely on word-boundary pattern matching. For Cyrillic languages such as Russian and Ukrainian, we also use stem-based matching to cover inflected forms. For French, we apply rule-based conjugation patterns, so verb forms beyond the infinitive are matched as well. For script-specific languages such as Arabic and Hebrew, we normalize text before searching for matches, whereas Chinese and Japanese rely on direct sub-string matching given the absence of word boundaries.

3.2.1. Zero-Shot Prompting

For zero-shot text detoxification experiments, we evaluate two widely used instruction-following LLMs, GPT-3.5-turbo (OpenAI, 2023) and GPT-4.1-mini (OpenAI, 2025), across all 15 languages using the templates presented in Table 3.

3.2.2. Fine-Tuning

Our framework explores two main sequence-to-sequence multilingual models: mT5 (Xue et al., 2021) and Aya-101 (Üstün et al., 2024). We select these models for three reasons. First, mT5 is the standard encoder-decoder baseline in the text detoxification literature (Dementieva et al., 2024a), ensuring direct comparability with prior work. Second, Aya-101 is one of the few open instruction-tuned multilingual models with broad language coverage, including low-resource languages, for which decoder-only alternatives such as LLaMA or Mistral offer more limited support. Third, comparing a base model (mT5) with an instruction-tuned model (Aya-101) allows us to isolate the effect of instruction tuning on the integration of toxic expressions in prompt instructions. This results in three model configurations: **mT5 (Full):** Full-parameter fine-tuning[9], **Aya-101 (Full):** Full-parameter fine-tuning[10], and **Aya-101 (LoRA):** Parameter-efficient fine-tuning via LoRA[11], enabling us to assess whether lightweight adaptation can match or surpass full fine-tuning in

[8] https://translate.google.com, accessed in November 2025.

[9] mT5 fine-tuned for 3 epochs with a learning rate of 3×10^{-5}, and early stopping with a patience of 4 evaluation steps.

[10] Aya-101 fine-tuned for 5 epochs with a learning rate of 2×10^{-4}, and early stopping with a patience of 5 evaluation steps.

[11] LoRA (Aya-101) fine-tuning using rank $r = 16$, $\alpha = 32$, trained for 5 epochs with a learning rate of 2×10^{-4}.

Model	BLEU	ROUGE	STA	SIM	CHRF	Joint
GPT-3.5-turbo w/o	17.40	19.94	0.59	0.56	0.31	0.15
GPT-3.5-turbo w/	27.24	25.08	0.75	0.72	0.52	0.32
GPT-4.1-mini w/o	25.22	25.54	**0.82**	0.74	0.50	0.33
GPT-4.1-mini w/	**38.96**	**30.93**	0.74	**0.85**	**0.63**	**0.41**

Table 4: Average performance of zero-shot prompting (GPT-3.5-turbo and GPT-4.1-mini) with and without toxic expressions in instructions across all 15 languages: *w/o*: instruction-only (Table 3), *w/*: toxic-expressions-in-instruction (Table 3).

this setting. We tune the models on each instruction strategy, resulting in six fine-tuned models in total.

All models are fine-tuned on the training data described in Section 3.1 and evaluated across all 15 languages (9 seen and 6 unseen), with temperature-based sampling ($\tau = 0.7$) to handle language imbalance.

3.3. Evaluation

We evaluate the detoxification performance by directly adapting three core metrics from Dementieva et al. (2024a,b). These metrics return values between 0 and 1 where higher is better.

Style Transfer Accuracy (STA) (Prabhumoye et al., 2018) is computed using a pretrained multilingual toxicity classifier[12] to score both the input and detoxified output, where a higher STA indicates more successful transfer from toxic to non-toxic style.

Content Similarity (SIM) is calculated using cosine similarity between the embeddings of the original toxic text and the generated detoxified text (Feng et al., 2022). It measures how well the original text's meaning is preserved in the output.

Fluency (CHRF) evaluates the fluency of the generated output. For this, an implementation from the *sacrebleu* library is used (Post, 2018).

Joint score (J) is the average of the product of STA, SIM and CHRF, which has been shown to be highly correlated with human evaluation (Logacheva et al., 2022).

Additionally, in line with other studies on text style transfer (Mukherjee et al., 2023b; Jin et al., 2022), we automatically evaluate model outputs using BLEU score (Papineni et al., 2002) and ROUGE[13] score (Lin and Och, 2004).

[12] https://huggingface.co/textdetox/xlmr-large-toxicity-classifier-v2

[13] ROUGE as the mean of ROUGE-1, ROUGE-2, and ROUGE-L

Model	BLEU	ROUGE	STA	SIM	CHRF	Joint
mT5 w/o	58.85	31.55	0.52	0.83	0.58	0.26
mT5 w/	**61.62**	33.05	0.56	0.90	0.66	0.35
Aya-LoRA w/o	49.83	33.93	0.66	0.89	0.69	0.42
Aya-LoRA w/	50.24	**33.99**	**0.71**	0.91	**0.70**	**0.45**
Aya-Full w/o	49.69	33.71	0.63	0.91	0.69	0.41
Aya-Full w/	51.71	33.96	0.70	**0.92**	0.69	0.44

Table 5: Average performance with and without toxic expressions in instructions across all 15 languages. *Aya-LoRA*: LoRA fine-tuning; *Aya-Full*: full-parameter fine-tuning: *w/o*: instruction-only (Table 3), *w/*: toxic-expressions-in-instruction (Table 3).

4. Results

In this section, we present the results from zero-shot prompting and fine-tuning experiments. We then discuss cross-lingual transfer results for the fine-tuned models across seen and unseen languages.

4.1. Zero-Shot Prompting Results

Table 4 presents the average performance of both GPT models across 15 languages. With instruction-only (w/o), GPT-3.5-turbo achieves a Joint score of 0.15, reflecting limited detoxification ability in a purely zero-shot setting, while GPT-4.1-mini performs considerably better at 0.33, likely benefiting from its more recent and capable pretraining. Adding toxic expressions in the instruction (w/) yields substantial gains over the standard strategy (w/o) in both cases. For GPT-3.5-turbo, the Joint score more than doubles from 0.15 to 0.32, with BLEU improving from 17.40 to 27.24. GPT-4.1-mini similarly benefits from the context, with the Joint score rising from 0.33 to 0.41 and BLEU from 25.22 to 38.96. This consistent pattern suggests that, without explicit guidance, it is not obvious to the models what constitutes toxic language based on their pretraining alone—likely due to the filtered nature of their training data (Mendu et al., 2025). Explicitly providing toxic expressions in instructions serves as a strong guiding signal, reducing ambiguity about what the model should remove or replace. A detailed per-language Joint score breakdown is provided in Table 6. Overall, models perform best on high-resource languages such as English, German, French, and Italian under both instruction settings, with a few exceptions, for example, GPT-3.5-turbo w/o performs poorly on Ukrainian. The toxic-expressions-in-instruction strategy brings the most dramatic improvements for GPT-3.5-turbo on these low-scoring languages, with Ukrainian jumping from 0.03 to 0.51 and Hindi from 0.01 to 0.23. GPT-4.1-mini shows more consistent performance across languages, though it similarly benefits from toxic expression guidance, particularly for

Model	EN	RU	UK	DE	AR	ES	HI	ZH	AM	FR	IT	JA	Hin	TT	HE	Avg.
GPT-3.5 w/o	0.37	0.27	0.03	0.34	0.17	0.19	0.01	0.02	0.02	0.33	0.21	0.08	0.04	0.02	0.02	**0.15**
GPT-3.5 w/	0.38	0.47	0.51	0.41	0.46	0.41	0.23	0.16	0.03	0.45	0.49	0.43	0.08	0.06	0.20	**0.32**
GPT-4.1-mini w/o	0.52	0.46	0.54	0.35	0.25	0.45	0.25	0.11	0.22	0.55	0.55	0.28	0.08	0.13	0.20	**0.33**
GPT-4.1-mini w/	**0.53**	**0.54**	**0.62**	**0.55**	**0.48**	**0.51**	**0.28**	**0.21**	**0.25**	**0.61**	**0.64**	**0.46**	**0.11**	**0.17**	**0.23**	**0.41**

Table 6: Joint score for zero-shot prompting across all 15 languages. *w/o*: instruction-only (Table 3); *w/*: toxic-expressions-in-instruction (Table 3). Language codes: EN: English, RU: Russian, UK: Ukrainian, DE: German, AR: Arabic, ES: Spanish, HI: Hindi, ZH: Chinese, AM: Amharic, FR: French, IT: Italian, JA: Japanese, Hin: Hinglish (Hindi in Latin script), TT: Tatar, HE: Hebrew.

	Seen Languages										Unseen Languages						
Model	EN*	RU*	UK*	DE*	AR*	ES*	HI*	ZH*	AM*	Avg.*	FR†	IT†	JA†	Hin†	TT†	HE†	Avg.†
mT5 w/o	0.40	0.37	0.38	0.41	0.39	0.31	0.16	0.07	0.19	**0.30**	0.31	0.34	0.24	0.10	0.12	0.13	**0.21**
mT5 w/	0.52	0.50	0.57	0.52	0.51	0.44	0.22	0.10	0.25	**0.40**	0.42	0.48	0.25	0.11	0.19	0.16	**0.27**
Aya-LoRA w/o	0.55	0.57	0.65	0.59	0.56	0.48	**0.26**	0.11	0.33	**0.46**	0.61	0.62	0.37	0.13	0.25	0.22	**0.37**
Aya-LoRA w/	**0.57**	**0.59**	**0.68**	**0.63**	**0.58**	**0.50**	0.24	**0.13**	**0.37**	**0.48**	**0.65**	**0.63**	**0.41**	**0.15**	**0.30**	**0.26**	**0.40**
Aya-Full w/o	0.54	0.56	0.65	0.60	0.55	0.45	0.24	0.11	0.32	**0.45**	0.57	0.58	0.35	0.13	0.26	0.22	**0.35**
Aya-Full w/	0.55	0.58	0.66	0.62	**0.58**	0.49	**0.26**	**0.13**	0.32	**0.47**	0.63	0.62	0.39	0.15	**0.30**	**0.26**	**0.39**

Table 7: Joint score for fine-tuned models across all 15 languages. $\star$: seen languages; languages were used for fine-tuning. †: unseen languages; not included during fine-tuning. *w/o*: instruction-only (Table 3); *w/*: toxic-expressions-in-instruction (Table 3). *Aya-LoRA*: LoRA fine-tuning; *Aya-Full*: full-parameter fine-tuning. Language codes: EN: English, RU: Russian, UK: Ukrainian, DE: German, AR: Arabic, ES: Spanish, HI: Hindi, ZH: Chinese, AM: Amharic, FR: French, IT: Italian, JA: Japanese, Hin: Hinglish (Hindi in Latin script), TT: Tatar, HE: Hebrew.

Japanese (0.28 $\rightarrow$ 0.46) and French (0.55 $\rightarrow$ 0.61). Chinese (ZH) and Hinglish (Hin) remain the most challenging languages for both models across both settings.

4.2. Fine-Tuning Results

Table 5 presents the average performance of all three fine-tuned models on 15 languages. mT5 w/o achieves a Joint score of 0.26, while both Aya variants perform considerably better at 0.42 (Aya-LoRA w/o) and 0.41 (Aya-Full w/o), reflecting the benefit of instruction-tuned pretraining. Adding toxic expressions in instructions consistently improves performance across all three models. For mT5, adding toxic expressions in instructions yields a noticeable gain in the Joint score (0.26 $\rightarrow$ 0.35), alongside improvements in BLEU, SIM, and CHRF. A similar trend is observed for both Aya-101 variants, where the toxic-expressions-in-instruction prompt improves the Joint score from 0.42 to 0.45 for Aya-LoRA and from 0.41 to 0.44 for Aya-Full. These findings confirm that providing toxic expressions in instructions consistently benefits model performance across all three model configurations, consistent with findings in the zero-shot experiments. Notably, Aya-LoRA outperforms Aya-Full in the instruction-enhanced setting across most metrics, achieving the highest Joint score (0.45) and the best STA and CHRF scores overall. This suggests that parameter-efficient fine-tuning via LoRA not only reduces computational cost but also yields a more effective model for the detoxification task despite using significantly fewer trainable parameters. A detailed breakdown of Joint score across all 15 languages, including performance on unseen languages, is provided in Table 7 and discussed further in Section 4.3.

4.3. Cross-Lingual Transfer Results

Table 7 presents the Joint score of fine-tuned models across seen ($\star$) and unseen (†) languages, revealing two consistent patterns. First, including toxic expressions in the prompt generally improves performance across nearly all settings. The gains vary by model but remain consistent overall, although Aya-LoRA and Aya-Full are exceptions in that they already achieve strong baseline performance with instruction-only prompts. Second, the fine-tuned models generalize remarkably well to unseen languages, with Aya-LoRA still achieving an average Joint score of 0.37–0.40 on unseen languages compared to 0.46–0.48 on seen ones—a marginal performance drop, yet still indicative of meaningful cross-lingual transfer. Looking at individual languages, French and Italian show the strongest performance among unseen languages (0.65 and 0.63 respectively on Joint score) even matching or exceeding the performance of some seen languages. This benefit is likely due to their typological proximity to Spanish and English, which were seen during training. In contrast, lower-resource and more distant languages such as Tatar, Hebrew, and Hinglish show considerably weaker results, suggesting that cross-lingual transfer is largely driven by language family similarity and script overlap.

Overall, Chinese (ZH) and Hinglish (Hin) remain the most challenging languages across all models

Original (Toxic)	Reference	Model Output	Model	Joint
*he is a f*cking parasite.*	*He is a bad person*	*he is a bad person.*	mT5 w/	0.0002
*you are a bit of a d*ck.*	*You are a bit intense*	*you are a bad person*	mT5 w/	0.0001
planet earth is fucking crazy.	*Planet earth is crazy*	*Planet earth is crazy.*	Aya-Full w/	0.0014
they re both fucked up.	*they are both in trouble*	*They are both messed up.*	Aya-Full w/	0.0003
*even bush knows he f*cked up.*	*Bush is well aware that he has messed up*	*Even Bush knows he messed up.*	Aya-LoRA w/	0.0002
liar, liar pants on fire justin.	*Constantly lying Justin*	*justin is lying.*	Aya-LoRA w/	0.0003
*dmvs are un f*cking believable.*	*dmvs are unbelievable*	*dmvs are unbelievable.*	GPT-4.1-mini w/	0.0006
*what an a*se hole you are!*	*What a bad person you are.*	*what a terrible person you are!*	GPT-4.1-mini w/	0.0002

Table 8: Examples of model outputs with low Joint score that may still be considered acceptable detoxifications, illustrating the limitations of automatic evaluation for this task.

and settings. For Chinese, the difficulty likely stems from its logographic script and limited overlap with other languages, which may also affect the quality of toxic span lookup. For Hinglish, the challenge is compounded by the unusual nature of the language itself, Hindi written in Latin script, which may not be well-represented in the pretraining data of any of the models, and whose toxic expressions may not align well with the entries in the toxic expressions lexicon where spelling may vary.

5. Conclusions and Discussion

In this work, we addressed the task of multilingual text detoxification with the objective of automatically transforming toxic text into a non-toxic rewrite without compromising meaning or fluency, making use of comparable toxicity lexicons. Our experiments were structured around two central questions: whether providing toxic expressions in prompt instructions improves detoxification performance, and whether the benefits of toxic expressions in instructions persist in a cross-lingual setup, including for unseen low-resource languages.

With respect to the first question, our results consistently confirm that providing toxic expressions in instructions yields measurable improvements across all models and settings, both in fine-tuning and zero-shot prompting. This holds for mT5, both variants of Aya-101, and the two GPT models, with the effect being particularly pronounced in zero-shot scenarios where no task-specific fine-tuning is available. These findings suggest that explicitly identifying and supplying toxic expressions reduces ambiguity for the model and serves as a reliable guiding signal for the detoxification process. Notably, we observe that a small set of high-frequency profane terms (e.g., *f*ck* in *English*, *p*tain* in *French*, *m*erda* in *Spanish* dominates the matches, while the long tail of the lexicon consists of entries that occur rarely or not at all.

With respect to the second question, the fine-tuned models demonstrate a notable ability to generalize to unseen languages, with a marginal yet meaningful drop in Joint scores compared to seen languages. Aya-LoRA, in particular, still achieves strong performance across both seen and unseen language settings, suggesting that parameter-efficient fine-tuning on multilingual data provides a robust foundation for cross-lingual transfer, even for low-resource languages such as Tatar. Overall, text detoxification remains challenging, particularly when trying to achieve detoxification while preserving content and maintaining fluency at the same time. Future work should incorporate human evaluation, more diverse prompting strategies, and additional low-resource languages.

More broadly, our results underscore the value of multilingual toxic expressions lexicons as a practical, transferable resource for injecting domain-specific knowledge into language models across languages and paradigms.

6. Limitations

One limitation of our work is the reliance on automatic evaluation metrics, which may not fully capture the quality of detoxification outputs. Given the sensitivity of this task to word choice and the degree of toxicity, small lexical changes can disproportionately affect scores such as BLEU or Joint, even when the output is semantically well-detoxified. As illustrated in Table 8, some outputs that received low automatic scores were nonetheless reasonable detoxified sentences upon inspection, highlighting the importance of human evaluation as a complementary assessment. A promising alternative would be LLM-as-a-judge evaluation, which we leave for future work. Furthermore, our detoxification task includes only explicitly toxic comments,

which themselves may have limited coverage, and does not address comments that convey implicit or inherently toxic messages. An example of the latter would be *"f*ck her right in the p*ssy"*, which carries an inherently toxic meaning that persists even when individual words are replaced. Paraphrasing such comments poses a challenge, as thorough detoxification may require altering or removing the original toxic meaning altogether (Dementieva et al., 2024a; Wiegand et al., 2023). Additionally, the use of machine-translated prompt templates across the 15 languages may introduce translation errors, particularly for low-resource languages such as Tatar and Amharic, which could affect model performance independently of the toxic expressions in instructions. Finally, while we experiment with two multilingual models, we acknowledge that larger model sizes and broader language coverage could yield further improvements (Ruan et al., 2024; Kaplan et al., 2020; Brown et al., 2020), which we leave for future work.

7. Ethical Considerations

The aim of our work is to detoxify toxic comments. However, what counts as toxicity is, to some extent, subjective. Automatic detoxification of user-generated comments in online environments should therefore be approached with caution. First, our models cannot guarantee the complete removal of toxicity with 100% accuracy. Second, automatic detoxification might be considered as a violation of freedom of speech (Dementieva et al., 2023, 2025). In line with the propositions made by Dementieva et al. (2025), our intention is to use the detoxification model for the creation of safer online environments and the reduction of harmful content and digital violence—not by enforcing automatic corrections, but rather by offering user-friendly suggestions for rephrasing potentially toxic messages. Finally, although our work is intended to be used for detoxification purposes, we cannot rule out the possibility of misuse, such as generating toxic text from non-toxic inputs (Bose et al., 2023; Floto et al., 2023).

8. Acknowledgments

We would like to acknowledge the reviewers for their helpful comments. We acknowledge the support of the Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg (MWK, Ministry of Science, Research and the Arts Baden-Württemberg under Az. 33-7533-9 19/54/5) in Künstliche Intelligenz & Gesellschaft: Reflecting Intelligent Systems for Diversity, Demography and Democracy (IRIS3D) and the support by the Interchange Forum for Reflecting on Intelligent Systems (IRIS) at the University of Stuttgart. We thank the Institute for Natural Language Processing (IMS), University of Stuttgart, for providing the computational resources used in this work.

9. Bibliographical References

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*, pages 6063–6074, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Ritwik Bose, Ian Perera, and Bonnie Dorr. 2023. Detoxifying online discourse: A guided response generation approach for reducing toxicity in user-generated text. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 9–14, Toronto, Canada. Association for Computational Linguistics.

Peter J Breckheimer. 2001. A haven for hate: The foreign and domestic implications of protecting internet hate speech under the first amendment. *S. Cal. L. Rev.*, 75:1493.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jennifer Cobbe. 2021. Algorithmic censorship by social platforms: Power and resistance. *Philosophy & Technology*, 34(4):739–766.

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*, pages 7979–7996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Trung Dang and Ferdinando D'Elia. 2025. Gemdetox at textdetox clef 2025: Enhancing a massively multilingual model for text detoxification on low-resource languages.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *International Conference on Web and Social Media*.

Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. 2024a. MultiParaDetox: Extending text detoxification with parallel data to new languages. In *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, Mexico City, Mexico. Association for Computational Linguistics.

Daryna Dementieva, Nikolay Babakov, Amit Ronen, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Daniil Moskovskiy, Elisei Stakovskii, Eran Kaufman, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2025. Multilingual and explainable text detoxification with parallel corpora. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7998–8025, Abu Dhabi, UAE. Association for Computational Linguistics.

Dementieva, Daryna and Moskovskiy, Daniil and Babakov, Nikolay and Ayele, Abinew Ali and Rizwan, Naquee and Schneider, Frolian and Wang, Xintong and Yimam, Seid Muhie and Ustalov, Dmitry and Stakovskii, Elisei and Smirnova, Alisa and Elnagar, Ashraf and Mukherjee, Animesh and Panchenko, Alexander. 2024b. *Overview of the Multilingual Text Detoxification Task at PAN 2024*. CEUR-WS.org.

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)*, page 1083–1101. Association for Computational Linguistics.

Daryna Dementieva, Sergey Ustyantsev, David Dale, Olga Kozlova, Nikita Semenov, Alexander Panchenko, and Varvara Logacheva. 2021. Crowdsourcing of parallel corpora: the case of style transfer for detoxification. In *Proceedings of the 2nd Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale co-located with 47th International Conference on Very Large Data Bases (VLDB 2021 (https://vldb.org/2021/))*, pages 35–49, Copenhagen, Denmark. CEUR Workshop Proceedings.

Ashwin Geet D'Sa, Irina Illina, and Dominique Fohr. 2020. Towards non-toxic landscapes: Automatic toxic comment detection using DNN. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 21–25, Marseille, France. European Language Resources Association (ELRA).

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)*, pages 878–891, Dublin, Ireland. 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.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Jared Kaplan, Sam McCandlish, Thomas Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *ArXiv*, abs/2001.08361.

Nicole Lai-Lopez, Lusha Wang, Su Yuan, and Liza Zhang. 2025. ylmmcl at multilingual text detoxification 2025: Lexicon-guided detoxification and classifier-gated rewriting.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*, ACL '04, pages 605–es. Association for Computational Linguistics.

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)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

Sai Krishna Mendu, Harish Yenala, Aditi Gulati, Shanu Kumar, and Parag Agrawal. 2025. Towards safer pretraining: Analyzing and filtering harmful content in webscale datasets for responsible llms. In *International Joint Conference on Artificial Intelligence*. Microsoft.

Sourabrata Mukherjee, Akanksha Bansal, Atul Kr. Ojha, John P. McCrae, and Ondrej Dusek. 2023a. Text detoxification as style transfer in English and Hindi. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 133–144, Goa University, Goa, India. NLP Association of India (NLPAI).

Sourabrata Mukherjee, Vojtěch Hudeček, and Ondřej Dušek. 2023b. Polite chatbot: A text style transfer application. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 87–93, Dubrovnik, Croatia. 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.

Gopala Krishna Nuthakki, Lekkala Sai Teja, and Atul Mishra. 2025. Team detox at textdetox clef 2025: Multilingual text detoxification using llm. In *Notebook for the PAN Lab at CLEF 2025*, volume 4038 of *CEUR Workshop Proceedings*, Madrid, Spain. CEUR-WS.org.

OpenAI. 2023. GPT-3.5 Turbo (gpt-3.5-turbo) model documentation. https://developers.openai.com/api/docs/models/gpt-3.5-turbo. Accessed: January 2026.

OpenAI. 2025. GPT-4.1 mini (gpt-4.1-mini) model documentation. https://developers.openai.com/api/docs/models/gpt-4.1-mini. Accessed: January 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Bahar Radfar, K. Shivaram, and Aron Culotta. 2020. Characterizing variation in toxic language by social context. In *International Conference on Web and Social Media*.

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. 2024. Observational scaling laws and the predictability of language model performance. In *Advances in Neural Information Processing Systems*, volume 37, pages 15841–15892. Curran Associates, Inc.

Elisei Rykov, Konstantin Zaytsev, Ivan Anisimov, and Alexandr Voronin. 2024. Smurfcat at PAN 2024 textdetox: Alignment of multilingual transformers for text detoxification. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*, volume 3740 of *CEUR Workshop Proceedings*, pages 2866–2871. CEUR-WS.org.

Zheyuan Ryan Shi, Claire Wang, and Fei Fang. 2020. Artificial intelligence for social good: A survey. *ArXiv*, abs/2001.01818.

Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley,

Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini. 2021. Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 247–267.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

P. N. Vasist, Debashis Chatterjee, and Satish Krishnan. 2023. The polarizing impact of political disinformation and hate speech: A cross-country configural narrative. *Information Systems Frontiers*, pages 1–26. Epub ahead of print.

Jeremy Waldron. 2012. *The Harm in Hate Speech*. Harvard University Press.

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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. 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*, pages 483–498, Online. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Chiyu Zhang, Honglong Cai, Yuezhang Li, Yuexin Wu, Le Hou, and Muhammad Abdul-Mageed. 2024. Distilling text style transfer with self-explanation from LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 200–211, Mexico City, Mexico. Association for Computational Linguistics.

10. Language Resource References

Dementieva, Daryna and Protasov, Vitaly and Babakov, Nikolay and Rizwan, Naquee and Alimova, Ilseyar and Brune, Caroline and Konovalov, Vasily and Muti, Arianna and Liebeskind, Chaya and Litvak, Marina and Nozza, Debora and Shah Khan, Shehryaar and Takeshita, Sotaro and Vanetik, Natalia and Ayele, Abinew Ali and Schneider, Florian and Wang, Xintog and Yimam, Seid Muhie and Elnagar, Ashraf and Mukherjee, Animesh and Panchenko, Alexander. 2025. *Overview of the Multilingual Text Detoxification Task at PAN 2025*. CEUR-WS.org, CEUR Workshop Proceedings.