

Multilingual and Explainable Text Detoxification with Parallel Corpora

Daryna Dementieva¹, Nikolay Babakov²,
Amit Ronen³, Abinew Ali Ayele^{4,5}, Naqee Rizwan⁶, Florian Schneider⁴,
Xintong Wang⁴, Seid Muhie Yimam⁴, Daniil Moskovskiy^{8,9}, Elisei Stakovskii¹⁰,
Eran Kaufman³, Ashraf Elnagar⁷, Animesh Mukherjee⁶, Alexander Panchenko^{8,9}

¹Technical University of Munich, ²Universidade de Santiago de Compostela, ³Shenkar College,

⁴University of Hamburg, ⁵Bahir Dar University, ⁶IIT Kharagpur, ⁷University of Sharjah,

⁸Skoltech, ⁹AIRI, ¹⁰University of North Carolina at Chapel Hill

daryna.dementieva@tum.de, a.panchenko@skol.tech

Abstract

Even with various regulations in place across countries and social media platforms (Government of India, 2021; European Parliament and Council of the European Union, 2022), digital abusive speech remains a significant issue. One potential approach to address this challenge is automatic text detoxification, a text style transfer (TST) approach that transforms toxic language into a more neutral or non-toxic form. To date, the availability of parallel corpora for the text detoxification task (Logacheva et al., 2022; Atwell et al., 2022; Dementieva et al., 2024a) has proven to be crucial for state-of-the-art approaches. With this work, we extend parallel text detoxification corpus to new languages—German, Chinese, Arabic, Hindi, and Amharic—testing in the extensive multilingual setup TST baselines. Next, we conduct the first of its kind an automated, explainable analysis of the descriptive features of both toxic and non-toxic sentences, diving deeply into the nuances, similarities, and differences of toxicity and detoxification across 9 languages. Finally, based on the obtained insights, we experiment with a novel text detoxification method inspired by the Chain-of-Thoughts reasoning approach, enhancing the prompting process through clustering on relevant descriptive attributes.

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

1 Introduction

The issue of managing toxic speech remains a crucial aspect of human communication and **digital violence** prevention (Shi et al., 2020), including the mitigation of toxic responses generated by Large Language Models (LLMs) (Yao et al., 2023). The typical approach to dealing with abusive speech on social platforms involves message blocking (Cobbe, 2021). To address this, numerous toxic and hate speech detection models have been developed for different languages, i.e. English (Mathew et al., 2021), Spanish (Molero et al.,

Language	Toxic Text	Detoxified Text
English	What a f**k is this about?	What is this about?
German	Was für ein besch**senes Jahr	Was für ein schlechtes Jahr.
Hindi	ये माद**द डरे हुए लग रहे हैं ?	ये लोग डरे हुए लग रहे हैं ?
Amharic	አንተ ቆሻሻ በዚህ ወቅት አይገህገ ማየት አልፈልግም	አንተ ጥሩ ሰው አይደለህም በዚህ ወቅት አንተን ማየት አልፈልግም
Arabic	تقتلوا القتل وتمشوا بجنائزته يا شر**ط	تقتلوا القتل وتمشوا بجنائزته
Chinese	卧槽，抓到了！	天啊，抓到了！

Figure 1: Examples of the desired texts detoxification for English and new languages: German, Chinese, Arabic, Hindi, and Amharic.

2023), Amharic (Ayele et al., 2023), Code-Mixed Hindi (Bohra et al., 2018), and many others (Costajussà et al., 2024). However, the recent research indicates a necessity for more proactive moderation of abusive speech (Kulenović, 2023). One such approach is **text detoxification**.

Within the baselines approaches for automatic text detoxification, multiple unsupervised baselines were created based on ideas of Delete-Retrieve-Generate (Li et al., 2018), latent style spaces disentanglement (Nogueira dos Santos et al., 2018), or conditional generation with Masked Language Modeling (Dale et al., 2021). However, the latest state-of-the-art outcomes, particularly in English, were attained when parallel data and fine-tuning with text-to-text generation models were employed as in ParaDetox (Logacheva et al., 2022) or APPDIA (Atwell et al., 2022). Then, several works were conducted to explore the potential

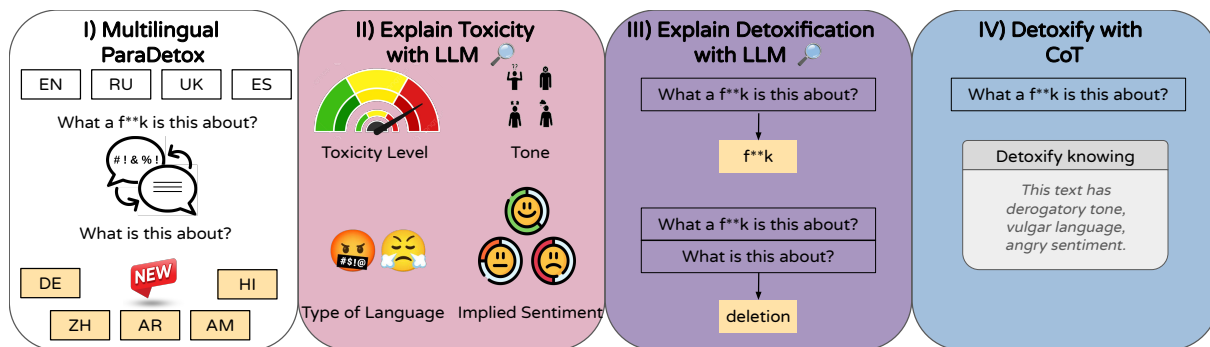


Figure 2: In this work, we extend parallel text detoxification data to new languages as well as provide explainability analysis of toxicity and detoxification attributes across all languages. This information helps to improve Chain-of-Thoughts reasoning for automatic text detoxification with LLMs.

of multilingual and cross-lingual text detoxification (Moskovskiy et al., 2022; Dementieva et al., 2024a). With this work, we extend the parallel text detoxification corpora to even more languages. Also, we are the first to conduct a comprehensive analysis of the full parallel multilingual corpus, uncovering unique traits and commonalities in how toxicity manifests across different languages and the ways to rephrase them. Thus, our contributions are the following (see Figure 2):

- We extend parallel text detoxification data to new languages—German, Chinese, Arabic, Hindi, and Amharic—thoroughly reporting each annotation process (Figure 2, I);
- We perform the first-of-its-kind study on explainability of parallel detoxification data thoroughly examining toxicity (Figure 2, II) and detoxification attributes (Figure 2, III) across 9 languages;
- Finally, we benchmark text detoxification baselines across a comprehensive multilingual dataset, incorporating a novel Chain-of-Thoughts prompting approach for detoxification with LLMs (Figure 2, IV).

All data, code, and analysis results are publicly accessible online.^{1,2,3}

2 Related Work

Modern Text Style Transfer Text style transfer (TST) methods can generally be categorized

¹<https://huggingface.co/textdetox>

²https://github.com/textdetox/multilingual_explainable_paradetox

³The data introduced in this work served as the foundation for the TextDetox CLEF-2024 Shared Task (Dementieva et al., 2024b).

into unsupervised and supervised approaches (Jin et al., 2022). Typically, when a text classification corpus for a specific domain is available, unsupervised methods are employed. For instance, cond-BERT and ParaGedi were introduced for controllable masked language modeling in (Dale et al., 2021), with MaRCO further enhancing these methods by incorporating multiple experts (Hallinan et al., 2023). Additionally, diffusion models have been explored for controllable text generation, particularly for text detoxification (Floto et al., 2023; Horvitz et al., 2024). Large Language Models (LLMs) have also shown promising results across various NLP tasks, including paraphrasing, leading to their application in different TST tasks (Mukherjee et al., 2024b), and specifically in text detoxification through the CoTex pipeline (Zhang et al., 2024). However, the availability of parallel training corpora has been shown to significantly enhance the performance of TST methods, often surpassing LLMs, which can be prone to hallucination. Such parallel corpora, though, are limited to specific tasks, including Bible historical styles (Carlson et al., 2018), GYAFC for formality (Rao and Tetreault, 2018), and APPDIA (Atwell et al., 2022) and ParaDetox (Logacheva et al., 2022) for detoxification.

Multilingual Text Style Transfer To date, several studies have explored text style transfer across various languages, extending beyond just English. For instance, sentiment transfer has been developed for Bangla (Mukherjee et al., 2023) and other Indian languages (Mukherjee et al., 2024a). In terms of formality, the English-focused GYAFC dataset was expanded to the X-FORMAL dataset (Briakou et al., 2021), which includes Brazilian Portuguese, French, and Italian. More recently, formality style

transfer has been examined for Japanese (Ung, 2023). Detoxification techniques have been applied to English (Logacheva et al., 2022), then Russian, Ukrainian, and Spanish (Dementieva et al., 2024a). However, these studies still primarily focus on European languages, leaving many other regions of the world unexplored.

Explainable Abusive Speech Mitigation To build trustworthy systems for mitigating different kinds of abusive speech, the aspect of explainability has gained increasing attention recently (Gongane et al., 2024). One of the first work in this area (Mathew et al., 2021) introduced the HateXplaine dataset, where annotators not only labeled the data but also provided the rationale behind their classifications. Following this, explainable AI frameworks like SHAP (Lundberg and Lee, 2017) and LIME (Ribeiro et al., 2016) have been applied to various text classification tasks, including hate and toxic speech (Mosca et al., 2023; Imbwaga et al., 2024). For toxic language specifically, the ToXCL framework (Hoang et al., 2024) was developed to fine-tune multiple models addressing different aspects of toxic speech detection. Additionally, recent advancements in LLMs have been leveraged for both text style transfer and generating corresponding explanations in the context of text detoxification (Khondaker et al., 2024).

3 New ParaDetox Annotation

We manually collected new data following the main quality criteria (Logacheva et al., 2022): (i) new paraphrases should be non-toxic; (ii) maximal content preservation; (iii) fluency on par with the original text. These data cover **five languages**—German, Hindi, Amharic, Arabic, and Chinese—chosen based on the native languages of the authors. Annotation and quality control were conducted either by the authors themselves or by hired assistants fluent in the respective languages.

Definition of Toxicity We adopt the definition introduced by Dementieva et al. (2024a) only addressing **vulgar or profane language** (Costa-jussà et al., 2022; Logacheva et al., 2022) while the overall message can be either toxic or neutral, but it should not involve deep insults or hate towards individuals or groups of people.

Data Preprocessing For all languages, we maintain the length of samples as sentences of around

Language	# Annot.	# Annot. Per Sent.	# Toxic Sent.	# Detoxified
German	2	1	3 521	1 103
Hindi	2	1	2 328	1 007
Amharic	2	1	2 995	1 000
Arabic	3	3	2 100	1 181
Chinese	3	1	1 380	1 000

Table 1: Summary of annotators and detoxifiable sentences statistics per language.

5-20 tokens. Also, if a text sample is from a social network, we anonymize or fully eliminate any mentioning of usernames and links.

Annotation Guidelines Firstly, we organized a joint meeting with all language stakeholders to present the base annotation guidelines in English. If needed, each language stakeholder adapted these guidelines to their own language and cultural context; otherwise, they reused the English version. All final annotation guidelines are made publicly available online.⁴

Annotators Compensations Compensation varied according to each university’s and country’s regulations. For German and Chinese, annotators were employed in Germany at a rate of €20 per hour (€7.65 above the minimum wage). For Amharic, the annotators were hired from Ethiopia \$5.8 per hour which is better than an M.Sc holder salary in the country. For Arabic and Hindi data, the annotators were existing lab or research projects employees receiving standard academic salaries.

Annotators Well-Being The annotation process took about four months providing flexible schedules and regular check-ins. Language stakeholders and task experts met weekly to discuss issues; daily stand-ups with annotators ensured supportive progress. Annotators could pause at any time without meeting daily quotas. Their expertise suited the project’s needs, and limitations emerged.

3.1 German

German ParaDetox was collected with several annotators with manual quality verification:

3.1.1 Input Data Preparation

The German language source data is based on three datasets containing toxic, offensive, or hate speech comments on social media about primarily political

⁴https://github.com/textdetox/multilingual_explainable_paradetox/tree/main/paradetox_collection

events in Germany or the US. For the two datasets from the GermEval 2018 (Wiegand et al., 2018) and GermEval 2021 (Risch et al., 2021) shared tasks, we used data from both the test and the train split. For the GermEval 2018 data, we only used samples labeled with the coarse class “*OFFENSE*” whereas for the GermEval 2021 data we only used samples annotated with the “*Subl_Toxic*” class. The third dataset (Ross et al., 2016) was filtered so only samples were kept where both expert annotators classified the samples as hate speech. The data from the three datasets was merged and deduplicated via exact string matching. As a result, 3 521 toxic were selected as candidates from which 1 103 were possible to detoxify.

3.1.2 Annotation Process

To create the final parallel detoxified German dataset, we hired two native German annotators. Annotator A is a female born in 1994 who holds a Master of Arts degree in Social Sciences, and Annotator B is a male born in 1992 who holds a Master of Science degree in Computer Science. The data was distributed so that each sample was transcribed by only one of the annotators.

3.2 Hindi

Hindi dataset was collected manually by native-speakers gaining data from multiple sources:

3.2.1 Input Data Preparation

We used the HASOC dataset created at FIRE 2019 (Mandl et al., 2019) as source for Hindi language. Contents in this dataset are relevant within Indian subcontinent which are collected from various social media platforms prevalent in India. For curation, posts containing *OFFENSIVE* and *PROFANE* contents in train and test splits were used. 1 455 *PROFANE* posts (1 237 train + 218 test) and 873 *OFFENSIVE* posts (676 train + 197 test) were chosen to prepare detoxifiable toxic data for our task. On a total of 2 328 samples, we first performed deduplication via exact string matching.

3.2.2 Annotation Process

Annotation Setup Out of 2328 samples, 1007 samples were marked as detoxifiable. Annotators were guided to re-write toxic pairs in a non-toxic manner, keeping the meaning of the original post unchanged.

Annotators One male NLP researcher working in the field of hate/toxic speech and another female

student enrolled in Bachelor’s Degree and having experience in Machine Learning, were employed to carry out the annotation. Both annotators are Indian, native Hindi speakers and are well versed with the topicality covered in the dataset. Each sentence was assigned to a single annotator. Afterwards, the data were cross-verified by a language stakeholder and domain experts.

3.3 Amharic

We compiled new Amharic ParaDetox datasets with the following annotation details:

3.3.1 Input Data Preparation

The input toxicity data is entirely sourced from the two previous studies, namely (Ayele et al., 2023) and (Ayele et al., 2022). We extracted a subset of these datasets labeled as *offensive*.

3.3.2 Annotation Process

Annotation Setup We customized the Potato-Portable Text Annotation Tool⁵ and utilized it for the annotation of Amharic ParaDetox dataset. Annotators were provided annotation guidelines, took hands-on practical training, completed independent training tasks before the main annotation task.

We began with a pilot annotation of 125 items by three native Amharic speakers and reviewed the quality in a group meeting with experts to clarify the task. Next, we annotated 2 995 tweets, each by a single annotator. Each tweet was classified as either detoxifiable or non-detoxifiable. Detoxifiable tweets were then rewritten in a detoxified manner.

Annotators Two annotators (one male and one female) were evolved in the main annotation, where both of them are university lecturers and have basic knowledge of NLP tasks.

3.4 Arabic

Here are details of Arabic ParaDetox collection:

3.4.1 Input Data Preparation

The Arabic ParaDetox dataset was created by combining parts of several existing datasets along with the Arabic-translated version of the Jigsaw dataset (Jigsaw, 2017). It includes the Levantine Twitter Dataset for Hate Speech and Abusive Language (L-HSAB) (Mulki et al., 2019), which focuses on Levantine dialects, and the Tunisian Hate and Abusive Speech (T-HSAB) dataset (Haddad et al., 2019), which targets Tunisian dialects. It also

⁵<https://github.com/davidjurgens/potato>

Language	Source of Toxic Samples	Annotation Process	Train	Test
English	(Jigsaw, 2017)	Crowdsourcing	400	600
Russian	(Belchikov, 2019; Semiletov, 2020)	CrowdSourcing	400	600
Ukrainian	(Bobrovnyk, 2019a)	Crowdsourcing	400	600
Spanish	(Pereira-Kohatsu et al., 2019; Taulé et al., 2024) (Pérez et al., 2022)	Crowdsourcing	400	600
German	(Wiegand et al., 2018; Risch et al., 2021) (Ross et al., 2016)	Manual	400	600
Hindi	(Mandl et al., 2019)	Manual	400	600
Amharic	(Ayele et al., 2023, 2022)	Manual	400	600
Arabic	(Mulki et al., 2019; Haddad et al., 2019) (Mubarak et al., 2020; Mulki and Ghanem, 2021)	Manual	400	600
Chinese	(Lu et al., 2023)	Manual	400	600

Table 2: All currently available ParaDetox datasets from previous work (Logacheva et al., 2022; Dementieva et al., 2024a) and the new ones. The human detoxified references were collected either via crowdsourcing or by hired native speakers. In this work, 1 000 samples per language were selected to perform analysis and experiments.

incorporates the OSACT dataset (Mubarak et al., 2020) and the Arabic Levantine Twitter Dataset for Misogynistic Language (LeT-Mi) (Mulki and Ghanem, 2021), which specifically addresses gender-based abuse. These resources combine to form the Arabic ParaDetox dataset, aimed at aiding the development of toxicity classifiers capable of handling Arabic content across various dialects and contexts. As a result, 2100 sentences were selected as candidates with 1181 were possible to detoxify.

3.4.2 Annotation Process

Annotators Detoxification was performed by three PhD-level annotators (two male, one female), all native Arabic speakers with strong computational linguistics backgrounds. Each text sample was transcribed by two annotators, and majority voting determined whether a sentence could be detoxified and if the resulting detoxification was appropriate.

3.5 Chinese

We collected new Chinese ParaDetox datasets with the following annotation details:

3.5.1 Input Data Preparation

Input Toxicity Data The Chinese ParaDetox dataset is derived from TOXICN (Lu et al., 2023), a recently released Chinese toxic language dataset. TOXICN was compiled from social media platforms and comprises 12 011 comments addressing several sensitive topics, including gender, race, region, and LGBTQ issues. From this dataset, we extracted a subset based on multiple criteria: the number of toxic words, the ratio of toxic words

in the comments, the length of comments, and the toxic scores of comments.

Input Preprocessing We set thresholds for the criteria: the number of toxic words ranged from 1 to 5 (checked by the predefined keywords list), the ratio of toxic words in comments was less than 0.5, and the length of comments ranged from 3 to 50 words, ensuring suitability for annotators to rewrite them. Subsequently, we employed a pre-trained toxic classifier (Lu et al., 2023) to compute the toxic scores of the selected comments, using a threshold score of 0.978 to filter the candidates. Ultimately, we collected 1 149 samples from the training set and 231 samples from the test set, resulting in a total of 1 380 samples deemed suitable for annotation.

3.5.2 Annotation Process

Annotation Setup For data annotation and verification, we employed a specifically designed three-task pipeline: *Task 1: Determine if the sentences are toxic.* Annotators were required to choose one of three options: the given sentence is *neutral*, *toxic but can be rewritten*, or *toxic and cannot be rewritten*. The last option was included based on the observation that some toxic texts are impossible to rewrite in a non-toxic manner. *Task 2: Rewrite sentences in a non-toxic style.* Annotators were instructed to create detoxified versions of the toxic sentences identified in Task 1 preserving the main content of the original sentences and rewriting the toxic words. *Task 3: Cross-check verification.* The detoxified sentences were assigned to different annotators for verification to ensure the quality.

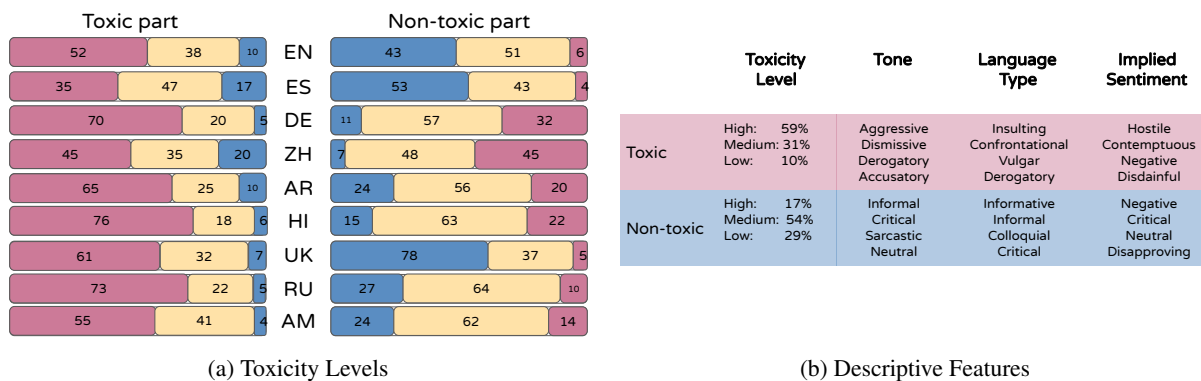


Figure 3: Extracted with GPT-4 toxicity levels and top descriptive features per toxic and non-toxic parts in the multilingual parallel text detoxification data.

Annotators We hired three native Chinese annotators from mainland China—two 22-year-old women with Bachelor’s degrees in engineering and one 32-year-old man with a Master’s in computer science—ensuring strong familiarity with both the language and the detoxification task.

3.6 Final Dataset

The full picture of newly collected and available for now parallel detoxification data in 9 languages is presented in Table 2. In the final stage, experts and native speakers thoroughly reviewed the entire dataset to ensure it met the task’s specific requirements and criteria. Using both existing (Logacheva et al., 2022; Dementieva et al., 2024a) and newly collected data, we selected 1 000 samples per language which were then split into 400 training and 600 test instances.⁶ These datasets and their respective divisions were subsequently utilized for further described analysis and experiments.

4 Explaining ParaDetox with LLM

Although Large Language Models (LLMs) still have room for improvement in text classification tasks, specifically, for hate and toxic speech (Roy et al., 2023), they have shown significant success in generating explanations (Singh et al., 2024). Given the resource-intensive nature of manually annotating descriptive aspects for each sample across multiple languages, we utilized GPT-4 to assist in generating explanations. We ensured the quality of these explanations by validating them with native speakers, while also conducting an in-depth analysis of parallel text detoxification data.

⁶https://huggingface.co/datasets/textdetox/multilingual_paradtox

4.1 Approach

For all our experiments, we employ GPT-4 (OpenAI, 2022) (May, 2024) leveraging the Chain-of-Thought reasoning method (Qiao et al., 2023) and the CO-STAR framework (Kwon and Gopalan, 2021) specifically designed for reasoning about toxicity and stereotypical biases in data to enhance the detoxification prompt design. All 1 000 pairs per nine languages were used for this analysis. The full texts of all prompts are available in Appendix A.

We compare toxic and detoxified parts to validate the detoxification process and identify cross-lingual similarities and differences in toxicity. For both parts, we extract descriptive features—toxicity level, tone, language type, implied sentiment, and negative connotation—using the following prompt (Appendix A.1, output example in Table 12):

Sentence: {sentence};

Toxicity Level: Specify here (Low/Medium/High);

Tone: the overall tone of the sentence—choose from keywords;

Language: Language style—choose from keywords;

Implied Sentiment: the overall sentiment—choose from keywords;

Context: Brief description of how context contributes to toxicity;

Negative Connotations: List specific negative words/phrases here.

We first prompted the model for open-ended descriptions for each feature, then selected the top 30 keywords from the explanations to refine the prompt, minimizing hallucinations. The core prompt was in English, with the target sentence in the respective language. Experts and native speakers reviewed all 1 000 samples per language for each feature and toxic keyword. All experts observed GPT-4’s tendency to overreact to certain

EN f*ck sh*t idiot d*ck as*	ES mierda (<i>sh*t</i>) subnormal (<i>subnormal</i>) puto (<i>f*cking</i>) culo (<i>a*s</i>) fascistas (<i>f*scists</i>)	DE arsch (<i>a*s</i>) dumm (<i>stupid</i>) lügenpresse (<i>lying</i> <i>press</i>) r*pefugees asylanten (<i>asylum</i> <i>seekers</i>)
ZH 恶心 (<i>disgusting</i>) 基佬 (<i>gay</i>) 舔狗 (<i>simp</i>) 普信男 (<i>average</i> <i>guy</i>) 垃圾 (<i>trash</i>)	AR ابن قحية (<i>son of a</i> <i>b*tch</i>) ادعسك (<i>smash you</i>) عبي (<i>idiot</i>) ايري (<i>c*cksucker</i>) خرا (<i>sh*t</i>)	HI भोसडी (<i>large c*nt</i>) भडवा (<i>pimps</i>) हरामी (<i>b*stard</i>) मादरचोद (<i>motherf*cker</i>) सूअर (<i>pigs</i>)
UK блядь (<i>f*ck</i>) хуй (<i>c*ck</i>) пиздец (<i>f*ck</i>) мудак (<i>a*shole</i>) йобаний (<i>f*cking</i>)	RU бля (<i>f*ck</i>) тварь (<i>creature</i>) хуй (<i>c*ck</i>) дебил (<i>m*ron</i>) пидор (<i>f*ggot</i>)	AM ደደብ (<i>dumb</i>) ደገቆሮ (<i>idiot</i>) ዑኸ (<i>dog</i>) ሰይጣን (<i>devil</i>) ቆሽሽ (<i>trash</i>)

Figure 4: Top-5 extracted keywords from toxic parts.

keywords, yet its toxicity rankings were accurate. For descriptive features and toxic keywords across all languages, experts agreed with GPT-4’s answers in 98% of cases.

4.2 Toxicity Descriptive Features Analysis

The overall view on top descriptive features for all languages as well as toxicity level per language are provided in Figure 3. The full list of top descriptive feature per language are provided in Appendix E.

Across all languages, we observe a reduction from high toxicity to medium or low levels, confirming that the paraphrases have been effectively detoxified. The original texts are predominantly *aggressive*, *derogatory*, *vulgar*, and *insulting*, often conveying *hostile*, *negative*, and *disdainful* sentiments. In contrast, the neutral paraphrases tend to shift towards *informal*, *colloquial*, or even *neutral* language, though they may still retain some *negative* or *critical* undertones.

4.3 Toxic Keywords Analysis

We extracted the most frequent toxic collocations from the toxic texts, as shown in Figure 4.

We found both similarities and differences in the typical rude and obscene language across languages. While some toxic words—like, *f*ck*, *idiot*, *as**—are present almost in all target languages, we can also see cultural specifics. In Ukrainian, Russian, and Chinese, derogatory comparisons involving homosexual individuals are considered insults, while in Hindi and Amharic, referring to someone using animal names is more prevalent. In Germany,

while the issue of temporarily displaced individuals sparks significant societal debate, rudeness often manifests through wordplay targeting these individuals. As a result, while common obscene language appears across all languages, the expressions of toxicity are culturally dependent thus requires culture-aware toxicity mitigation solutions.

4.4 Text Detoxification Analysis

Then, we analyzed the way how detoxification was performed (see Table 3). We sought lemmas that reflect various editorial actions—*delete*, *remove*, *rephrase*, *replace*, *insert*, *add*—using the following prompt template: Answer shortly, how this text: {toxic text} was rephrased into this: {detoxified text}. Additionally, we computed the Levenshtein distance between toxic and non-toxic parts (Appendix D).

Across all languages, adding new content is rare. Detoxification mainly involves removing or rephrasing toxic elements. In German, Arabic, Hindi, Ukrainian, and Amharic, removal and rephrasing occur equally, while Spanish favors removal and Chinese/Russian rely more on rephrasing. Consequently, localized edits with fluent substitutions generally suffice for effective detoxification.

Lang.	Del	Rep	Ins	Lang.	Del	Rep	Ins
EN	27%	60%	13%	HI	47%	44%	9%
ES	60%	26%	14%	UK	37%	59%	4%
DE	44%	48%	8%	RU	23%	71%	6%
ZH	14%	84%	2%	AM	45%	44%	11%
AR	35%	55%	10%				

Table 3: Percentage of toxic phrases **Deleted**, **Rephrased**, or new non-toxic parts **Inserted** in order to achieve detoxification.

4.5 Chain-of-Thoughts Text Detoxification

Finally, we developed a new chain-of-thought reasoning approach to improve text detoxification with LLMs by guiding detoxification with explanations and close examples.

Our descriptive analysis suggests that the most effective detoxification approach varies according to descriptive features, toxicity expression and the target language itself. Depending on these factors, the detoxification strategy should be chosen accordingly. While it is challenging to come up with the clear human-readable instruction, the detoxification can be explained via examples. Thus, based on the extracted descriptive features, we performed

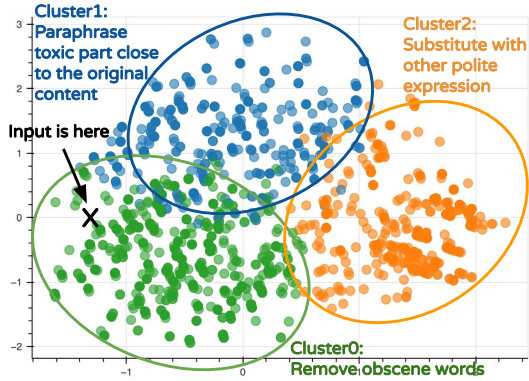


Figure 5: Text detoxification with CoT: analyze the input, identify its cluster, and provide the detoxification explanation and cluster example in the prompt.

K-means clustering per language on their one-hot encodings. The experiments with hyperparameters indicated an optimal division into 3 clusters with the following top descriptive features and approximate explanations (Figure 5, i.e. for English):

- **Cluster 0:** *Offensive, hostile*, and characterized by *vulgar* language. Texts can be detoxified mainly by *removing* profanities.
- **Cluster 1:** *Condescending, derogatory, dismissive*, and potentially *biased* by gender or race. Here, texts requires more significant *rephrasing* to remove condescending or biased language.
- **Cluster 2:** *Informal, casual, and playful*. Texts can be slightly adjusted by *inserting* neutral or polite expressions after removing the toxic parts.

Upon receiving new input, the LLM first estimates the descriptive features of a new text and the corresponding clustering is performed. LLM is then prompted to detoxify this sentence now using information about the cluster and a representative example of how to detoxify this type of cluster. The full prompt example can be found in Appendix A.3 and clusters details in Appendix F.

5 Automatic Evaluation Setup

We adopt the evaluation pipeline from Logacheva et al. (2022) to our multilingual setup. Direct links to the datasets/models instances are in Appendix B.

Style Transfer Accuracy (STA) We subsampled 5 000 samples—2 500 toxic and 2 500 neutral—from toxicity classification corpora for each language (see in Table 2) that were not used for ParaDetox data collection. We fine-tuned XLM-R-large (Conneau et al., 2020) instance for the binary toxicity classification task.

Content Similarity (SIM) is the cosine similarity between LaBSE embeddings (Feng et al., 2022) of the source texts and the generated texts.

Fluency (ChrF1) is used to estimate the proximity of the detoxified texts to human references. we use an implementation of ChrF1 score from sacrebleu library (Post, 2018).

Joint score (J) is the aggregation of the three above metrics:

$$J = \frac{1}{n} \sum_{i=1}^n \text{STA}(y_i) \cdot \text{SIM}(x_i, y_i) \cdot \text{ChrF1}(x_i, y_i),$$

where $\text{STA}(y_i)$, $\text{SIM}(x_i, y_i)$, $\text{ChrF1}(x_i, y_i) \in [0, 1]$ for each text detoxification output y_i .

6 Baselines

For comparison, we considered several unsupervised and supervised text detoxification approaches together with a baseline prompt construction. Details of the hyperparameters and model choices for each method can be found in Appendix C.

Duplicate Trivial baseline: the output sentence is a copy-paste of the input sentence. This baseline has 1.0 (or 100%) SIM score by definition.

Delete Removal of offensive terms using a manually compiled list of vulgar words. We collected and compiled together the lists of such toxic keywords for all target languages based on openly available sources (see Table 5).

Backtranslation As for a more sophisticated unsupervised baseline, we performed translation of non-English texts into English with NLLB (Costajussà et al., 2022) to then perform detoxification with the fine-tuned on English ParaDetox BART (Logacheva et al., 2022). The detoxification results were translated back to the target languages.

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.

Fine-tuned LM on Translated Data We also tried to obtain synthetic parallel corpora by translating selected 400 English ParaDetox samples to our target languages. We utilized mBART for machine translation model (Liu et al., 2020) for the translation step. We tuned the mBART for text generation (Tang et al., 2020) on the obtained data.

	Average	EN	ES	DE	ZH	AR	HI	UK	RU	AM
Human References	0.608	0.711	0.709	0.733	0.201	0.695	0.298	0.790	0.732	0.601
<i>Unsupervised Approaches</i>										
Duplicate	0.126	0.061	0.090	0.287	0.069	0.294	0.035	0.032	0.048	0.217
Delete	0.302	0.447	0.319	0.362	<u>0.175</u>	<u>0.456</u>	0.105	0.328	0.255	<u>0.270</u>
Backtranslation	0.205	<u>0.506</u>	0.275	0.233	0.027	0.206	0.104	0.201	0.223	0.075
condBERT	0.213	0.278	0.347	0.310	0.067	0.337	0.033	0.316	0.224	0.003
<i>Supervised Approaches</i>										
mBART-Translated	0.291	0.443	0.315	0.392	0.083	0.365	0.142	0.343	0.359	0.178
mBART-mParaDetox	0.282	0.339	0.289	<u>0.409</u>	0.068	0.397	0.171	0.345	0.321	0.204
<i>LLM-based Approaches</i>										
GPT-4 few-shot	0.324	0.475	0.422	0.396	0.109	0.270	0.194	0.460	0.383	0.205
GPT-4 CoT (ours)	<u>0.331</u>	0.326	<u>0.447</u>	0.400	0.117	0.339	0.251	<u>0.503</u>	<u>0.426</u>	0.166

Table 4: Results of the *automatic* evaluation of the text detoxification approaches. The scores for each language are respective **Joint** scores. **Bold** denote the best results within the group, **underlined**—the best for the language.

Fine-tuning on the parallel data Finally, we fine-tuned the multilingual text-to-text generation model mBART-Large on the selected training multilingual data.

GPT-4 few-shot prompting Before CoT, we applied a few-shot prompting of GPT-4 with the example prompt presented in Appendix A.2.

7 Results

We conducted a multilingual text detoxification across all languages on the test sets, with the results presented in Table 4 and detailed metrics per language in Appendix G. Surprisingly, the Delete method outperformed other unsupervised approaches for three languages—Chinese, Arabic, and Amharic. This may be due to the nature of these languages (Table 3), where detoxification relies heavily on paraphrasing. Since the proposed methods still struggled with appropriate paraphrasing, Delete, which removes toxic content without rephrasing, performed best. However, for other languages, where rephrasing is also key, LM-based solutions excelled, likely due to better representation of the languages in the pre-training data.

While for the majority of languages mBART fine-tuned on human-curated data outperformed the model fine-tuned on translated data, this results is not consistent. As described previously, some obscene terms are similar across languages and can be translated from English, offering sufficient information about toxicity for the target language. However, in the case of German, Hindi, Ukrainian,

and Amharic cultural nuances play a significant role, leading the model trained on manually crafted data to perform better.

Finally, incorporating cluster information into the prompting process significantly boosted GPT-4 CoT’s performance, surpassing the few-shot prompting approach for nearly all languages. This suggests that targeting toxicity with greater precision and information on relevant human-curated detoxifications reduces model hallucinations. As a result, this method achieved the highest scores across all approaches in the STA metric and standouts with the highest average J score (see example in Table 11).

8 Conclusion

This work addressed the multilingual and explainability aspects of the text detoxification task. We introduced manually curated parallel detoxification datasets for new languages—German, Chinese, Arabic, Hindi, and Amharic—and the detailed data collection process. Next, we used LLMs as explainability tools on nine languages to analyze key descriptive features of toxic and non-toxic texts, identify top toxic collocations, and determine the primary actions required for detoxification per different toxicity expressions. Building on these insights, we developed a new Chain-of-Thoughts LLM prompting text detoxification method that incorporates detoxification cluster information about the input text. This approach reduced model’s hallucinations, improved precision in edits, incorporated cultural specifics, and outperformed all baselines.

Limitations

Firstly, while the work aims to extend data to new languages, there remains significant room for improvement in incorporating as many languages as possible. The selection of languages in this study was based on the native languages of the authors, but broader involvement of other language stakeholders could enhance the dataset.

Secondly, this work focuses solely on multilingual detoxification without exploring monolingual or cross-lingual tasks. Further research could be conducted to identify the most effective detoxification model for each language using the created data. Additionally, cross-lingual approaches could explore how detoxification knowledge transfers between languages, opening new avenues for research. Preliminary cross-lingual transfer experiments have been conducted for English and Russian (Dementieva et al., 2023), but the new dataset now includes more languages for further exploration.

For the CoT approach, we focused on human-readable cluster explanations in English; however, this approximation was not thoroughly explored for other languages. Our method currently relies on example-based explanations, and further research into human-readable cluster descriptions remains open for future work.

Lastly, the primary experiments in this study were conducted using GPT-4, a closed-source model from OpenAI. While GPT-4 continues to perform exceptionally well in various NLP benchmarks, demonstrating stable generation of coherent explanations, we recognize the importance of supporting open-source initiatives. Therefore, we acknowledge the necessity of ablation study with opensource LLMs.

Ethics Statement

We explore the task of text detoxification with no intent to violate the freedom of speech, but rather to help mitigate digital violence, create safer online environments for children, and promote the development of secure AI models. The ideal implementation of detoxification models on communication platforms would be as suggestions, rather than forced corrections. A user-friendly interface for these suggestions should be considered by stakeholders.

Additionally, detoxifying LLMs, not just human content, is a relevant topic. Already several ap-

proaches were explored (Leong et al., 2023; Wang et al., 2024) utilizing English ParaDetox data as instruction dataset to mitigate toxicity in the model. However, these efforts have been limited to monolingual contexts due to data constraints. Further research into detoxifying LLMs in other languages, as well as the potential for cross-lingual knowledge transfer, represents a promising area for future study.

Finally, the authors of this work utilized ChatGPT to check the grammar and correct the appropriateness of the used language.

Acknowledgements

This work was only possible to the massive support from various institutions. Firstly, DD, NR, SMY, DM, AM and AP would like to thank SPARC-II (Scheme for Promotion of Academic and Research Collaboration, Phase II) project for funding international travel and subsistence to carry out this work. Then, pilot experiments and additional annotation was supported by Toloka.ai research grant. The further contribution of DD of this work was supported by the Friedrich Schiedel Fellowship hosted by the TUM School of Social Sciences and Technology and the TUM Think Tank. We sincerely acknowledge the financial support provided by the fellowship. Additionally, we would like to extend our gratitude to the TUM Data Analytics&Statistics chair, under the leadership of Alexander Fraser.

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, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 6063–6074. International Committee on Computational Linguistics.
- Abinew Ali Ayele, Skadi Dinter, Tadesse Destaw Belay, Tesfa Tegegne Asfaw, Seid Muhie Yimam, and Chris Biemann. 2022. *The 5Js in Ethiopia: Amharic hate speech data annotation using Toloka Crowdsourcing Platform*. In *Proceedings of the 4th International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 114–120, Bahir Dar, Ethiopia.
- Abinew Ali Ayele, Seid Muhie Yimam, Tadesse Destaw Belay, Tesfa Asfaw, and Chris Biemann. 2023. *Exploring Amharic hate speech data collection and classification approaches*. In *Proceedings of the 14th*

- International Conference on Recent Advances in Natural Language Processing*, pages 49–59, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- 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. The dictionary of ukrainian obscene words. <https://github.com/saganoren/obscene-ukr>. Accessed: 2024-12-12.
- Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media*, pages 36–41, New Orleans, Louisiana, USA. Association for Computational Linguistics.
- 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.
- Jennifer Cobbe. 2021. Algorithmic censorship by social platforms: Power and resistance. *Philosophy & Technology*, 34(4):739–766.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.
- Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.
- Marta R. Costa-jussà, Mariano Coria Meglioli, Pierre Andrews, David Dale, Prangthip Hansanti, Elahe Kalbassi, Alexandre Mourachko, Christophe Ropers, and Carleigh Wood. 2024. Mutox: Universal multilingual audio-based toxicity dataset and zero-shot detector. *CoRR*, abs/2401.05060.
- 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.
- 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, Varvara Logacheva, Irina Nikishina, Alena Fenogenova, David Dale, I. Krotova, Nikita Semenov, Tatiana Shavrina, and Alexander Panchenko. 2022. RUSSE-2022: Findings of the First Russian Detoxification Shared Task Based on Parallel Corpora. *COMPUTATIONAL LINGUISTICS AND INTELLECTUAL TECHNOLOGIES*.
- Daryna Dementieva, Daniil Moskovskiy, Nikolay Babakov, Abinew Ali Ayele, Naqee Rizwan, Florian Schneider, Xintog Wang, Seid Muhie Yimam, Dmitry Ustalov, Elisei Stakovskii, Alisa Smirnova, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2024b. Overview of the multilingual text detoxification task at pan 2024. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*. 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)*, pages 1083–1101, Nusa Dua, Bali. Association for Computational Linguistics.
- 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.
- European Parliament and Council of the European Union. 2022. [Regulation \(eu\) 2022/2065 of the european parliament and of the council of 19 october 2022 on a single market for digital services \(digital services act\) and amending directive 2000/31/ec](#). Official Journal of the European Union, L 277, 27.10.2022, p. 1–102.
- 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.
- Robert James Gabriel. 2023. English full list of bad words and top swear words banned by google. <https://github.com/coffee-and-fun/google-profanity-words/blob/main/data/en.txt>. Accessed: 2024-12-12.
- Vaishali U. Gongane, Mousami V. Munot, and Alwin D. Anuse. 2024. [A survey of explainable AI techniques for detection of fake news and hate speech on social media platforms](#). *J. Comput. Soc. Sci.*, 7(1):587–623.
- Government of India. 2021. [Information technology \(intermediary guidelines and digital media ethics code\) rules, 2021](#). Ministry of Electronics and Information Technology, Government of India.
- Hatem Haddad, Hala Mulki, and Asma Oueslati. 2019. [T-HSAB: A tunisian hate speech and abusive dataset](#). In *Arabic Language Processing: From Theory to Practice - 7th International Conference, ICALP 2019, Nancy, France, October 16-17, 2019, Proceedings*, volume 1108 of *Communications in Computer and Information Science*, pages 251–263. Springer.
- 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)*, pages 228–242, Toronto, Canada. Association for Computational Linguistics.
- Nhat Hoang, Xuan Long Do, Duc Anh Do, Duc Anh Vu, and Anh Tuan Luu. 2024. [ToXCL: A unified framework for toxic speech detection and explanation](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6460–6472, Mexico City, Mexico. Association for Computational Linguistics.
- Zachary Horvitz, Ajay Patel, Chris Callison-Burch, Zhou Yu, and Kathleen R. McKeown. 2024. [Paraguide: Guided diffusion paraphrasers for plug-and-play textual style transfer](#). In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18216–18224. AAAI Press.
- Joan L. Imbwaga, Nagaratna B. Chittaragi, and Shashidhar G. Koolagudi. 2024. [Explainable hate speech detection using LIME](#). *Int. J. Speech Technol.*, 27(3):793–815.
- Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. [SWSR: A chinese dataset and lexicon for online sexism detection](#). *Online Soc. Networks Media*, 27:100182.
- Jigsaw. 2017. Toxic comment classification challenge. <https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge>. Accessed: 2024-03-18.
- 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.
- Md Tawkat Islam Khondaker, Muhammad Abdul-Mageed, and Laks V. S. Lakshmanan. 2024. [Greenlama: A framework for detoxification with explanations](#). *CoRR*, abs/2402.15951.
- Enes Kulenović. 2023. [Should democracies ban hate speech? hate speech laws and counterspeech](#). *Ethical Theory and Moral Practice*, 26(4):511–532.
- Teyun Kwon and Anandha Gopalan. 2021. [CO-STAR: conceptualisation of stereotypes for analysis and reasoning](#). *CoRR*, abs/2112.00819.
- Chak Tou 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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4433–4449. Association for Computational Linguistics.
- Juncen Li, Robin Jia, He He, and Percy Liang. 2018. [Delete, retrieve, generate: a simple approach to sentiment and style transfer](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1865–1874. 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)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.
- Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, and Hongfei Lin. 2023. [Facilitating fine-grained detection of Chinese toxic language: Hierarchical taxonomy, resources, and benchmarks](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 16235–16250.
- Scott M. Lundberg and Su-In Lee. 2017. [A unified approach to interpreting model predictions](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774.
- Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. [Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages](#). In *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '19*, page 14–17, New York, NY, USA. Association for Computing Machinery.
- Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. [Hatexplain: A benchmark dataset for explainable hate speech detection](#). In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14867–14875. AAAI Press.
- José María Molero, Jorge Pérez-Martín, Álvaro Rodrigo, and Anselmo Peñas. 2023. [Offensive language detection in spanish social media: Testing from bag-of-words to transformers models](#). *IEEE Access*, 11:95639–95652.
- Edoardo Mosca, Daryna Dementieva, Tohid Ebrahim Ajdari, Maximilian Kummeth, Kirill Gringauz, Yutong Zhou, and Georg Groh. 2023. [IFAN: An explainability-focused interaction framework for humans and NLP models](#). 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: System Demonstrations*, pages 59–76, Bali, Indonesia. Association for Computational Linguistics.
- Daniil Moskovskiy, Daryna Dementieva, and Alexander Panchenko. 2022. [Exploring cross-lingual text detoxification with large multilingual language models](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 346–354, Dublin, Ireland. Association for Computational Linguistics.
- Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020. [Overview of OSACT4 Arabic offensive language detection shared task](#). In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 48–52, Marseille, France. European Language Resource Association.
- 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.
- Sourabrata Mukherjee, Atul Kr. Ojha, Akanksha Bansal, Deepak Alok, John P. McCrae, and Ondřej Dusek. 2024a. [Multilingual text style transfer: Datasets & models for indian languages](#). *CoRR*, abs/2405.20805.
- Sourabrata Mukherjee, Atul Kr. Ojha, and Ondřej Dusek. 2024b. [Are large language models actually good at text style transfer?](#) *CoRR*, abs/2406.05885.
- Hala Mulki and Bilal Ghanem. 2021. [Let-mi: An Arabic Levantine Twitter dataset for misogynistic language](#). In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 154–163, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. [L-HSAB: A Levantine Twitter dataset for hate speech and abusive language](#). In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118, Florence, Italy. 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.
- OpenAI. 2022. [Chatgpt: Optimizing language models for dialogue](#). Accessed: 2024-05-31.
- 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.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.
- Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. [Reasoning with language model prompting: A survey](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.
- 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.
- Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ["why should I trust you?": Explaining the predictions of any classifier](#). In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM.
- Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. [Overview of the germeval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments](#). In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments, GermEval@KONVENS 2021, Düsseldorf, Germany, September 6, 2021*, pages 1–12. Association for Computational Linguistics.
- Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. [Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis](#). In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, volume 17 of *Bochumer Linguistische Arbeitsberichte*, pages 6–9, Bochum, Germany.
- Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. [Probing LLMs for hate speech detection: strengths and vulnerabilities](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128, Singapore. 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.
- Zheyuan Ryan Shi, Claire Wang, and Fei Fang. 2020. [Artificial intelligence for social good: A survey](#). *CoRR*, abs/2001.01818.
- Inc Shutterstock. 2020. [List of dirty, naughty, obscene, and otherwise bad words](#). <https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words>. Accessed: 2024-12-12.
- Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. [Rethinking interpretability in the era of large language models](#). *CoRR*, abs/2402.01761.
- Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Namian Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. [Multilingual translation with extensible multilingual pretraining and finetuning](#). *CoRR*, abs/2008.00401.
- Mariona Taulé, Montserrat Nofre, Víctor Bargiela, and Xavier Bonet Casals. 2024. [Newscom-tox: a corpus of comments on news articles annotated for toxicity in spanish](#). *Lang. Resour. Evaluation*, 58(4):1115–1155.
- Rachel Ung. 2023. [Formality Style Transfer between Japanese and English](#). Ph.D. thesis, Waseda University.
- Shang Wang, Tianqing Zhu, Bo Liu, Ming Ding, Xu Guo, Dayong Ye, Wanlei Zhou, and Philip S. Yu. 2024. [Unique security and privacy threats of large language model: A comprehensive survey](#). *CoRR*, abs/2406.07973.
- Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. [Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language](#). In

Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Eric Sun, and Yue Zhang. 2023. [A survey on large language model \(LLM\) security and privacy: The good, the bad, and the ugly](#). *CoRR*, abs/2312.02003.

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.

A Prompts for Explanations and Chain-of-Thoughts Detoxification with LLMs

Here, we provide exact prompts used for explaining multilingual parallel detoxification data and text detoxification prompting.

A.1 Prompt for Descriptive Features Extraction

Prompt Instructions

Please analyze the provided sentence using the structure below to identify elements of toxicity and suggest improvements, when I tell you, use words from the keywords list (can be more than one word!):

keywords = [Neutral, Informative, Casual, Assertive, Dismissive, Condescending, Friendly, Commanding, Instructive Derogatory, Confrontational, Insulting, Vulgar, Formal, Informal, Offensive, Technical, Playful, Positive, Frustration, Analytical, Professional, Hostile, Hatred, Helpful, Angry, Friendly, Arrogant]

Analysis Structure (do not use " and [] and "" in your answer and do not suggest improvement!):

```
{  
  Sentence: {sentence},  
  Toxicity Level: Specify here (Low/Medium/High),  
  Tone: the overall tone of the sentence- choose from keywords,  
  Language: Language style—choose from keywords,  
  Implied Sentiment: the overall sentiment- choose from keywords,  
  Context: Brief description of how context contributes to toxicity,  
  Negative Connotations: List specific negative words/phrases here,  
  Intent: Describe the perceived intent behind the sentence.  
}
```

A.2 Few-Shot Prompting for Text Detoxification

Prompt Instructions

Please detoxify the provided sentence using the structure below without changing the real meaning of the sentence.

Analysis Structure (do not use " and [] and "" in your answer and do not suggest improvement!):

```
{  
  Sentence: {sentence},  
  Fixed sentence: <the non-toxic sentence without changing the meaning>  
},  
Example 1: {  
  Sentence: dude should have been taken to api , he would be right at home with all the other  
  knuckleheads there,  
  Fixed sentence: It would have been good if he went to api. He would fit in.  
}  
Example 2: {  
  Sentence: damn those young mothers driving their children to daycare through the snow  
  drifts.,  
  Fixed sentence: those young mothers driving their children to daycare through the snow drifts.  
}
```


A.3 Chain-of-Thoughts Prompting with Cluster Knowledge Incorporation

Prompt Instructions

Please detoxify the provided sentence using the structure below without changing the real meaning of the sentence.

The sentences are **clustered into 3 groups** while each group has its own characterizes.

Cluster 0 is more Offensive, Hostile and Vulgar;

Cluster 1 is more Condescending, Derogatory and Hostile;

Cluster 2 is more Informal, Casual, Dismissive.

For each sentence and cluster that I give, **make the sentence non-toxic by making it Neutral/Informal/Casual without changing the meaning.**

Analysis Structure (do not use " and [] and "" in your answer and do not suggest improvement!):

```
{  
  Sentence: {sentence},  
  Toxicity level: {Specify here},  
  Cluster: {cluster},  
  Fixed sentence: <the non-toxic sentence after making it Neutral/Informal/Casual without  
changing the meaning>;  
},
```

Example:

```
{  
  Sentence: dude should have been taken to api , he would be right at home with all the other  
knuckleheads there,  
  Toxicity Level: Medium,  
  Cluster: 0,  
  Fixed sentence: It would have been good if he went to api. He would fit in.  
}
```

B Automated Evaluation Metrics Models

The direct links to the datasets and models instances used for the evaluation setup:

- The compiled toxicity binary classification dataset to fine-tune an STA classifier;⁷
- The fine-tuned XLM-RoBERTa for STA estimation;⁸
- LaBSE multilingual encoder for SIM metric.⁹

C Hyperparameters Configurations for Considered Text Detoxification Approaches

Here, we provide the final hyperparameters and other details for the main considered text detoxification baselines, fine-tuned multilingual text generation models, and GPT-4.

C.1 Delete

The resources used for the multilingual toxicity lexicon compilation are listed in Table 5. The full list is available online for public usage and reproducibility.¹⁰

Language	Original Source	# of Keywords
English	(Logacheva et al., 2022; Gabriel, 2023; Costa-jussà et al., 2022)	3 390
Spanish	(Costa-jussà et al., 2022)	1 200
German	(Shutterstock, 2020; Costa-jussà et al., 2022)	247
Chinese	(Jiang et al., 2022; Lu et al., 2023; Costa-jussà et al., 2022)	3 840
Arabic	Ours+(Costa-jussà et al., 2022)	430
Hindi	(Costa-jussà et al., 2022)	133
Ukrainian	(Bobrovnyk, 2019b; Costa-jussà et al., 2022)	7 360
Russian	(Dementieva et al., 2022; Costa-jussà et al., 2022)	141 000
Amharic	Ours+(Costa-jussà et al., 2022)	245

Table 5: The list of the original sources and the corresponding amount of obscene keywords used to compile multilingual toxic lexicon list for our Delete baseline.

C.2 Backtranslation

For the translation step, we used the NLLB instance.¹¹ For English sentences detoxification, we utilized previously released BART-detox English instance.¹²

C.3 condBERT

We re-used of the condBERT pipeline introduced in (Dale et al., 2021)¹³ with mBERT-base¹⁴ model and the hyperparameters for the masked language modelling task via MaskedTokenPredictorBert class with parameters max_len= 250 and contrast_penalty= 0.0.

C.4 mBART

Previous experiments in Dementieva et al. (2024a) showed quite poor performance of BloomZ-7b (Muennighoff et al., 2023) for the text detoxification. To choose the model for supervised fine-tuning for new multilingual text detoxification, we compared in this case two multilingual text generation models—mT0-large (Muennighoff et al., 2023)¹⁵ and mBART-large (Tang et al., 2020)¹⁶. The results comparison

⁷https://huggingface.co/datasets/textdetox/multilingual_toxicity_dataset

⁸<https://huggingface.co/textdetox/xlmr-large-toxicity-classifier>

⁹<https://huggingface.co/sentence-transformers/LaBSE>

¹⁰huggingface.co/datasets/textdetox/multilingual_toxic_lexicon

¹¹<https://huggingface.co/facebook/nllb-200-distilled-600M>

¹²<https://huggingface.co/s-nlp/bart-base-detox>

¹³https://github.com/s-nlp/detox/tree/main/emnlp2021/style_transfer/condBERT

¹⁴<https://huggingface.co/google-bert/bert-base-multilingual-cased>

¹⁵<https://huggingface.co/bigscience/mt0-xxl-mt>

¹⁶<https://huggingface.co/facebook/mbart-large-50>

based on the overall J scores per language is presented in Table 6. In the end, for the final results, we chose mBART fine-tuned with the following setup: num_train_epochs= 10, warmusteps= 10, learning_rate= $1e - 05$, batch_size= 32. For the inference, we used the default parameters of MBartForConditionalGeneration class: beams_number= 5, maximal_tokens= 200.

	Average	EN	ES	DE	ZH	AR	HI	UK	RU	AM
Human References	0.608	0.711	0.709	0.733	0.201	0.695	0.298	0.790	0.732	0.601
	<i>Supervised Approaches</i>									
mT0-Translated	0.261	0.467	0.341	0.356	0.073	0.331	0.106	0.254	0.283	0.142
mT0-mParaDetox	0.168	0.397	0.107	0.244	0.002	0.356	0.150	0.040	0.119	0.097
mBART-Translated	0.291	0.443	0.315	0.392	0.083	0.365	0.142	0.343	0.359	0.178
mBART-mParaDetox	0.282	0.339	0.289	0.409	0.068	0.397	0.171	0.345	0.321	0.204

Table 6: Results of the *automatic* evaluation of the text detoxification approaches. The scores for each language are respective Joint scores. **Bold** denote the best results within the group.

C.5 GPT-4 Prompting

We employed GPT-4 (OpenAI, 2022) for analysis and experiments during May, 2024. We used default hyperparameters for the inference step which included temperature= 1.0, top_p= 1.0, top_k= 0.0, frequency_penalty= 0.0, presence_penalty=0.0.

D Toxic and Detoxified Sentences Lengths Comparison

Additionally to the toxic keywords and edits types analysis, we also provide the lengths comparison of toxic and non-toxic parallel pairs (Figure 6) and the Levenshtein distances between them (Figure 7). The lengths and distances calculation are based on the tokenization performed with `textdetox/xlmr-large-toxicity-classifier` used for STA calculation. Here, we again observe language-specific differences. For instance, in Chinese, detoxified versions are longer than their toxic counterparts, while in Amharic the length disparity is substantial. Even though toxic phrases are removed, the size of the replacement phrases can vary depending on both the language and the nature of the toxicity.

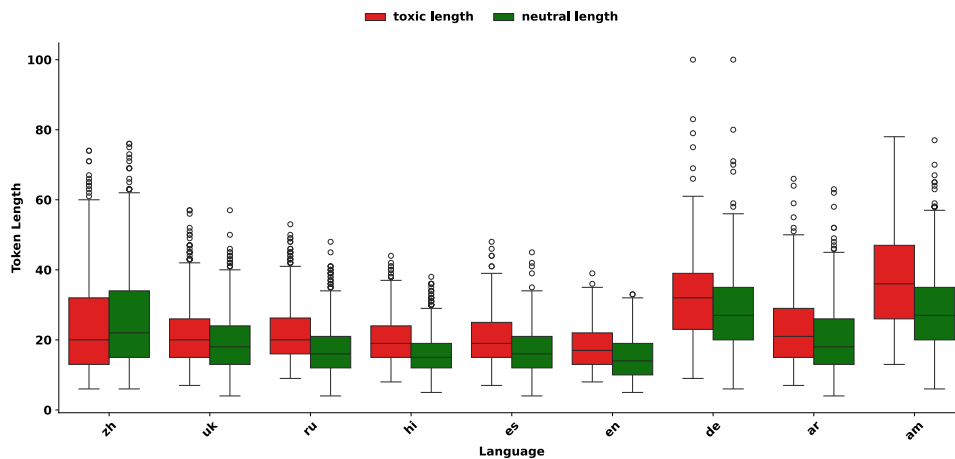


Figure 6: Comparison of toxic and non-toxic texts lengths distributions per each language.

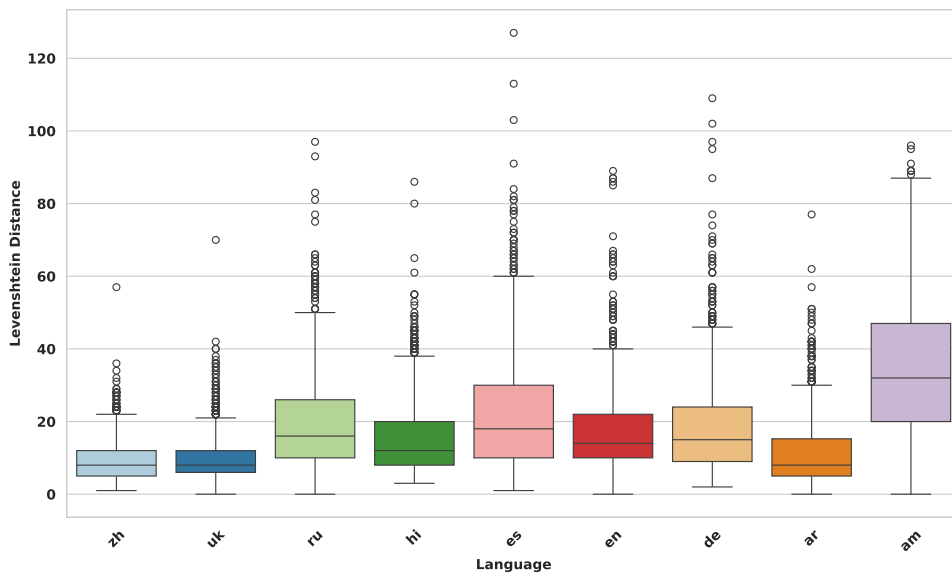
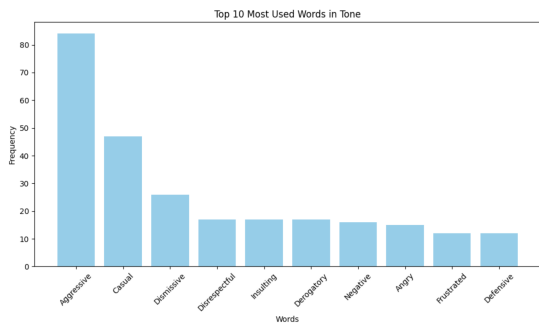


Figure 7: Levenshtein distances between toxic and non-toxic parts distribution.

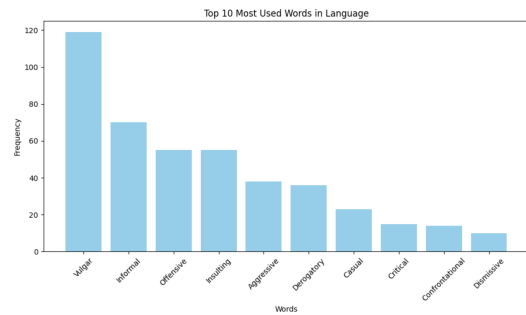
E Top Descriptive Features

Toxicity Level	Tone	Lang. Type	Implied Sentim.		Implied Sentim.	Lang. Type	Tone	Toxicity Level
High: 52% Med: 38% Low: 10%	Aggressive Frustrated Dismissive Derogatory	Vulgar Insulting Confrontat. Informal	Hostile Negative Angry Critical	EN	Negative Neutral Frustrat. Positive	Informal Informative Direct Critical	Informal Critical Neutral Accusatory	High: 6% Med: 51% Low: 43%
High: 35% Med: 47% Low: 17%	Aggressive Frustrated Dismissive Insulting	Vulgar Insulting Informal casual	Hostile Negative Contempt. Angry	ES	Negative Neutral Frustrat. Positive	Informal Informative Colloquial Neutral	Informal Neutral Sarcastic Critical	High: 4% Med: 43% Low: 53%
High: 70% Med: 25% Low: 5%	Aggressive Dismissive Derogatory Accusatory	Insulting Derogatory Confrontat. Offensive	Hostile Negative Angry Disdainful	DE	Negative Critical Disapprov. Disparaging	Informal Informative Colloquial Neutral	Informal Sarcastic Accusatory Critical	High: 32% Med: 57% Low: 11%
High: 45% Med: 35% Low: 20%	Dismissive Derogatory Aggressive Neutral	Insulting Derogatory Confrontat. Casual	Hostile Contempt. Negative Disdainful	ZH	Negative Critical Disapprov. Dismissive	Informative Informal Critical Derogatory	Informal Critical Sarcastic Neutral	High: 45% Med: 48% Low: 7%
High: 65% Med: 25% Low: 10%	Aggressive Insulting Dismissive Accusatory	Insulting Confrontat. Offensive Derogatory	Hostile Contempt. Negative Disrespect	AR	Negative Critical Neutral Hostile	Informative Critical Informal Colloquial	Critical Informal Accusatory Sarcastic	High: 20% Med: 56% Low: 24%
High: 76% Med: 18% Low: 6%	Aggressive Derogatory Insulting Accusatory	Insulting Offensive Derogatory Vulgar	Hostile Contempt. Disrespect Negative	HI	Negative Hostile Informal Aggressive	Informal Colloquial Critical Informative	Accusatory Critical Informal Aggressive	High: 22% Med: 63% Low: 15%
High: 61% Med: 32% Low: 7%	Aggressive Frustrated Dismissive Casual	Vulgar Insulting Confrontat. Offensive	Hostile Negative Angry Contempt.	UK	Negative Neutral Frustration Dismissive	Colloquial Informal Informative Conversat.	Informal Neutral Casual Sarcastic	High: 5% Med: 37% Low: 78%
High: 73% Med: 22% Low: 5%	Aggressive Dismissive Insulting Derogatory	Insulting Confrontat. Offensive Vulgar	Hostile Contempt. Negative Disdainful	RU	Negative Critical Neutral Disapprov.	Informative Colloquial Informal Critical	Informal Critical Accusatory Sarcastic	High: 9% Med: 64% Low: 27%
High: 55% Med: 41% Low: 4%	Aggressive Accusatory Derogatory Critical	Insulting Confrontat. Derogatory Critical	Hostile Contempt. Disapprov. Negative	AM	Negative Disapprov. Critical Neutral	Critical Informal Accusatory Confrontat.	Critical Accusatory Informal Confrontat.	High: 14% Med: 62% Low: 24%

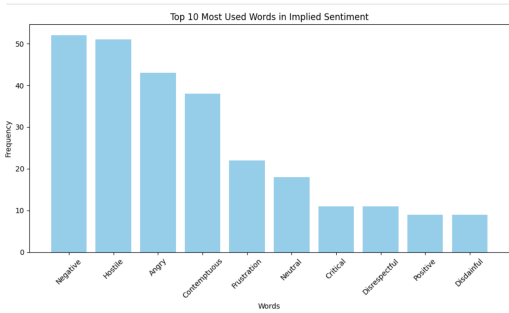
Table 7: Main descriptive features per language for **toxic** (on the left) and **detoxified** (on the right) parts.



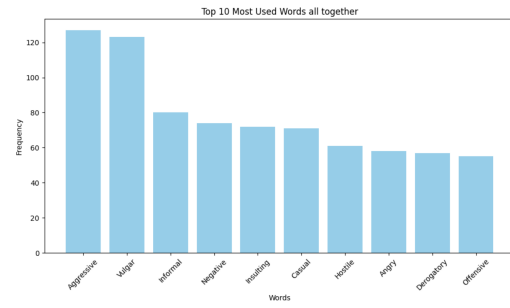
(a) Tone



(b) Language Type

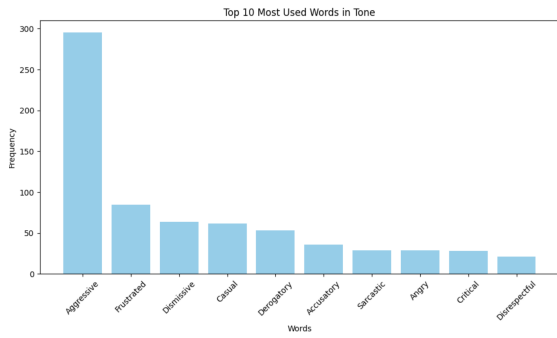


(c) Sentiment

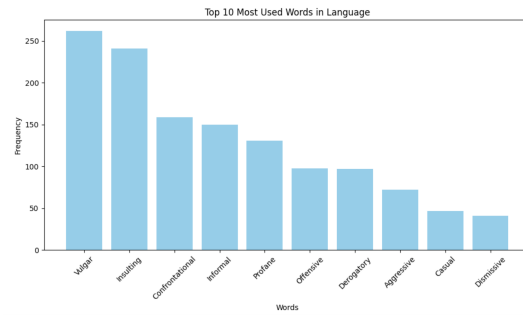


(d) All together

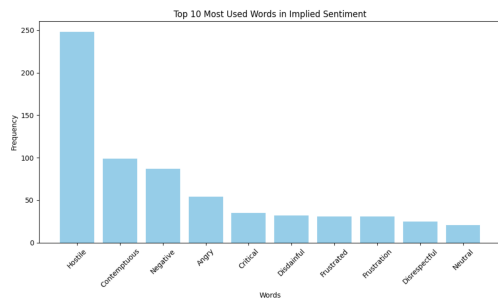
Figure 8: Descriptive words of the different features in the toxic training part for all languages.



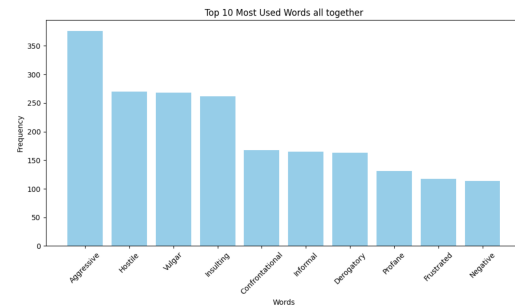
(a) Tone



(b) Language Type



(c) Sentiment



(d) All together

Figure 9: Descriptive words of the different features in the toxic test part for all languages.

F K-means Clustering Result Examples

Here, we present the 2D PCA projection of English toxic texts, one-hot-encoded with descriptive features, along with the resulting cluster divisions. (Figure 10).

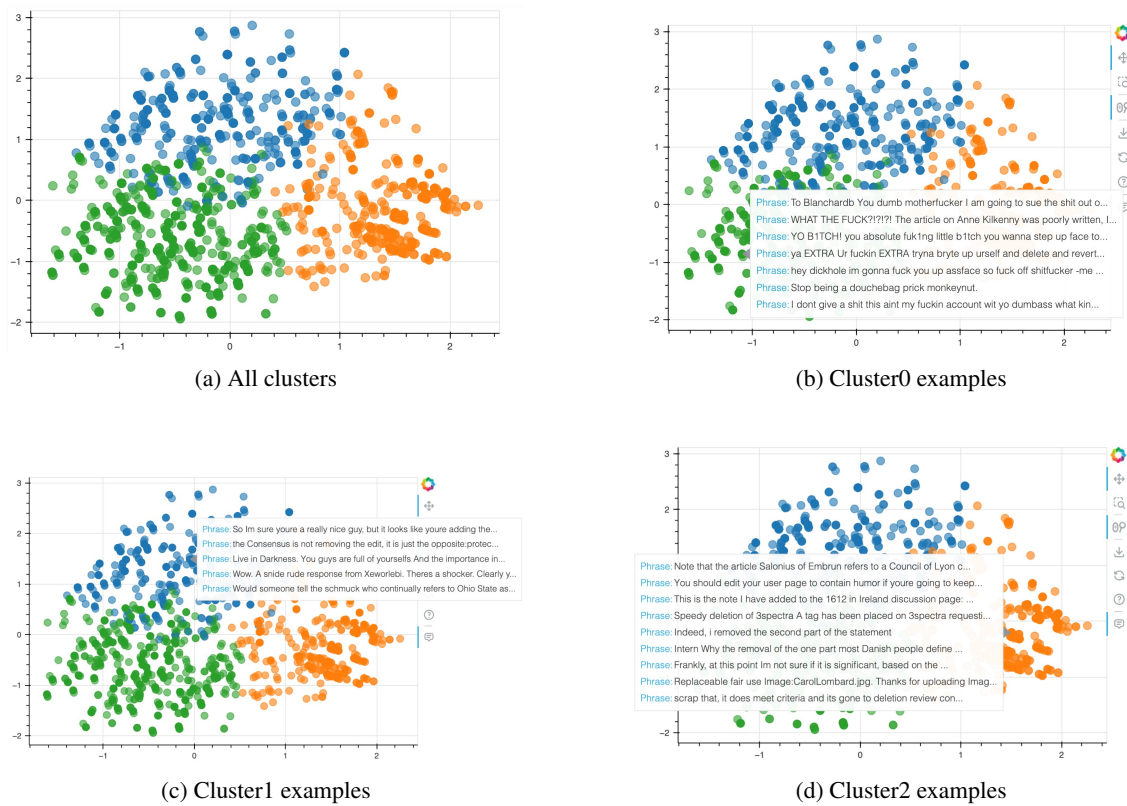


Figure 10: The PCA projection of the toxic sentences cluster based on their descriptive features and detoxification types.

G Automatic Evaluation Results per Language per Metric

Here, we provide the extended results of automatic evaluation setup based on all three evaluation parameters for all languages: English, Spanish, and German (Table 8); Chinese, Arabic, and Hindi (Table 9); Ukrainian, Russian, and Amharic (Table 10).

	English				Spanish				German			
	STA	SIM	ChrF	J	STA	SIM	ChrF	J	STA	SIM	ChrF	J
Human References	0.864	0.820	1.000	0.711	0.875	0.811	1.000	0.708	0.809	0.909	1.000	0.732
<i>Unsupervised Approaches</i>												
Duplicate	0.090	0.999	0.670	0.061	0.139	0.999	0.655	0.089	0.352	0.999	0.812	0.287
Delete	0.662	0.956	0.691	0.447	0.479	0.972	0.669	0.318	0.454	0.989	0.802	0.361
Backtranslation	0.807	0.868	0.693	0.506	0.812	0.770	0.423	0.275	0.796	0.747	0.372	0.232
condBERT	0.443	0.941	0.640	0.278	0.610	0.920	0.602	0.347	0.419	0.966	0.753	0.310
<i>Supervised Approaches</i>												
mBART-Translated	0.691	0.894	0.694	0.443	0.607	0.877	0.587	0.315	0.581	0.929	0.729	0.392
mBART-mParaDetox	0.493	0.934	0.695	0.339	0.474	0.933	0.635	0.289	0.532	0.969	0.794	0.409
<i>LLM-based Approaches</i>												
GPT-4 few-shot	0.807	0.865	0.661	0.475	0.867	0.806	0.584	0.421	0.683	0.888	0.659	0.395
GPT-4 CoT	0.985	0.682	0.454	0.326	0.949	0.789	0.573	0.447	0.908	0.783	0.544	0.400

Table 8: Automatic evaluation results for English, Spanish, and German. **Bold** denote the best results within the group, **underlined**—the best for the language.

	Chinese				Arabic				Hindi			
	STA	SIM	ChrF	J	STA	SIM	ChrF	J	STA	SIM	ChrF	J
Human References	0.266	0.789	1.000	0.201	0.795	0.875	1.000	0.694	0.367	0.814	1.000	0.297
<i>Unsupervised Approaches</i>												
Duplicate	0.130	0.999	0.535	0.069	0.388	0.999	0.776	0.293	0.051	0.999	0.695	0.034
Delete	0.384	0.887	0.524	0.174	0.597	0.974	0.777	0.455	0.146	0.974	0.706	0.104
Backtranslation	0.661	0.591	0.070	0.026	0.836	0.682	0.319	0.205	0.443	0.731	0.289	0.103
condBERT	0.138	0.993	0.518	0.067	0.488	0.957	0.726	0.337	0.050	0.976	0.667	0.033
<i>Supervised Approaches</i>												
mBART-Translated	0.272	0.901	0.356	0.083	0.626	0.899	0.667	0.365	0.243	0.896	0.617	0.142
mBART-mParaDetox	0.166	0.963	0.433	0.068	0.560	0.950	0.742	0.397	0.234	0.939	0.699	0.171
<i>LLM-based Approaches</i>												
GPT-4 few-shot	0.452	0.805	0.328	0.108	0.759	0.755	0.466	0.270	0.476	0.786	0.509	0.193
GPT-4 CoT	0.716	0.683	0.228	0.117	0.931	0.712	0.476	0.339	0.611	0.745	0.533	0.251

Table 9: Automatic evaluation results for Chinese, Arabic, and Hindi. **Bold** denote the best results within the group, **underlined**—the best for the language.

	Ukrainian				Russian				Amharic			
	STA	SIM	ChrF	J	STA	SIM	ChrF	J	STA	SIM	ChrF	J
Human References	0.877	0.899	1.000	0.790	0.887	0.824	1.000	0.732	0.893	0.683	1.000	0.601
<i>Unsupervised Approaches</i>												
Duplicate	0.037	0.999	0.778	0.031	0.067	0.999	0.698	0.048	0.426	0.999	0.485	0.216
Delete	0.423	<u>0.974</u>	<u>0.791</u>	<u>0.327</u>	0.372	<u>0.971</u>	<u>0.708</u>	<u>0.254</u>	0.539	<u>0.979</u>	<u>0.486</u>	<u>0.269</u>
Backtranslation	<u>0.914</u>	0.704	0.293	0.201	<u>0.903</u>	0.697	0.328	0.222	0.819	0.618	0.135	0.075
condBERT	0.424	0.960	0.759	0.316	<u>0.339</u>	0.944	0.666	0.224	<u>0.998</u>	0.169	0.007	0.003
<i>Supervised Approaches</i>												
mBART-Translated	<u>0.610</u>	0.870	0.647	0.343	<u>0.601</u>	0.885	0.657	<u>0.359</u>	0.501	<u>0.875</u>	0.391	0.178
mBART-mParaDetox	0.462	<u>0.939</u>	<u>0.751</u>	<u>0.345</u>	0.455	<u>0.937</u>	<u>0.709</u>	0.321	<u>0.506</u>	0.915	<u>0.412</u>	<u>0.204</u>
<i>LLM-based Approaches</i>												
GPT-4 few-shot	0.895	<u>0.842</u>	0.592	0.460	0.758	<u>0.825</u>	<u>0.600</u>	0.382	0.467	<u>0.946</u>	<u>0.453</u>	<u>0.205</u>
GPT-4 CoT	<u>0.955</u>	0.823	<u>0.612</u>	<u>0.503</u>	<u>0.939</u>	0.782	0.560	<u>0.426</u>	<u>0.835</u>	0.672	0.269	0.166

Table 10: Automatic evaluation results for Ukrainian, Russian, and Amharic. **Bold** denote the best results within the group, **underlined**—the best for the language.

Original	and nobody gave a sh*t .
Delete	and nobody gave a .
BART-detox (en)	and nobody cared.
CondBERT	and nobody gave a buttu .
mBART-Translated	and nobody gave a d*mn.
mBART-ParaDetox	and nobody gave a sh*t.
GPT-4 few-shot	and nobody cared.
GPT-4 CoT	and nobody seemed to care.
Original	everyone s got that single issue that represents the burr up their a*s
Delete	everyone s got that single issue that represents the burr up their .
BART-detox (en)	everyone s got that single issue that represent the burr up.
CondBERT	everyone s got that single issue that represents the burr up their Ass .
mBART-Translated	everyone s got that single issue that represents the burr up their a*s
mBART-ParaDetox	everyone s got that single issue that represents the burr up their a*s.
GPT-4 few-shot	everyone has that one issue that bothers them.
GPT-4 CoT	Everyone has that one issue that really bothers them.

Table 11: Examples of text detoxification outputs by different models for English for general readers to showcase the approached behaviour. For the phrases that require significant rephrasing, LLM, especially, with proposed CoT method suggests more reasonable detoxification. For mBART, it seems challenging to grasp detoxification knowledge properly for nine languages simultaneously.

H Multilingual ParaDetox Data Examples

Here, we provide an example with extracted features for English (Table 12) for general readers and several examples of data samples from new collected parallel text detoxification data for new languages: German (Table 13), Hindi (Table 11), Amharic (Table 12), Chinese (Table 13), and Arabic (Table 14).

Original	You made a mistake you a**.
Toxicity Level	High
Tone	Aggressive
Language	Insulting and offensive
Implied Sentiment	Hostile
Toxic Keywords	a**
Paraphrase	You made a mistake.
Toxicity Level	Low
Tone	Assertive
Language	Informal and straightforward
Implied Sentiment	Negative
Toxic Keywords	None
Original	holy sh*t i thought that was a joke .
Toxicity Level	Medium
Tone	Casual
Language	Informal
Implied Sentiment	Surprised
Toxic Keywords	sh*t
Paraphrase	I thought it was a joke!!
Toxicity Level	Low
Tone	Casual
Language	Informal
Implied Sentiment	Surprise or skepticism
Toxic Keywords	None

Table 12: Examples of parallel detoxified pairs from EnParaDetox.

Original	लेकिन यह भी साफ है रंडी कि धोनी कोहली से बेहतर है <i>But it is also clear you wh**e that Dhoni is better than Kohli.</i>
Paraphrase	लेकिन यह भी साफ है कि धोनी कोहली से बेहतर है <i>But it is also clear that Dhoni is better than Kohli.</i>
Original	किसी बहन के लौड़े की अवाज नहीं निकली <i>Not a single sister's penis spoke.</i>
Paraphrase	किसी की भी अवाज नहीं निकली <i>Not a single person spoke.</i>
Original	तेरी तरह थूक चाटने वाले कुत्ते नहीं हैं हम <i>We are not dogs that lick spit like you.</i>
Paraphrase	तेरी तरह किसी की चापलूसी नहीं करते हैं हम <i>We don't flatter anyone like you.</i>

Figure 11: Examples of parallel detoxified pairs from HiParaDetox.

Original	idi*****her Kommentar. Aufm Supermannheft gepennt? <i>Idi***c comment. Slipped up on the Superman magazine?</i>
Paraphrase	schlechter Kommentar. Aufm Supermannheft gepennt? <i>bad comment. Slipped up on the Superman magazine?</i>
Original	Ich will dieses A*****och nicht auf freiem Fuß wissen...egal in welchem Land..! <i>I don't want this a**ole at liberty...no matter in which country...!</i>
Paraphrase	Ich will diese Person nicht auf freiem Fuß wissen...egal in welchem Land..! <i>I don't want this person at liberty...no matter in which country...!</i>
Original	Ich finde er ist einfach ein unlustiger Spas**i <i>I just think he's an unfunny ret**d</i>
Paraphrase	Ich finde er ist einfach nicht lustig <i>I just don't think he's funny</i>

Table 13: Examples of parallel detoxified pairs from DeParaDetox.

Original	እረ እህቶቼ በዙ ሰዎች.?? ይቺ ፈንጋ እንኳን እንደ ወንድሜ እኮ ነው የማይህ ትበለኝ ?? <i>Oh people, we are getting more sisters?? This ugly girl even told me that she sees me as her brother??</i>
Paraphrase	እረ እህቶቼ በዙ ሰዎች.?? ይቺም እንደ ወንድሜ እኮ ነው የማይህ ትበለኝ ?? <i>Oh people, we are getting more sisters??, this girl told me that she sees me as her brother??</i>
Original	እንዳንተ አይነቱ ደደብ ቢያጠፋም ባያጠፋም በቀን አንድ መመከር እበት <i>A d*mp like you, whether he makes mistake or not, should be advised everyday</i>
Paraphrase	ለአንተ በቀን አንድ ምክር መሰጠት ሳያሰፈልግ አይቀርም <i>I think we need to give you advice everyday.</i>
Original	አሳረረን እኮ እግዚአብሔር ይድፋው እሄ እርኩስ ደግሞ አይመረጥም <i>He has embittered us, may God kill him, this filthy one will never be elected</i>
Paraphrase	አስመረረን እኮ እግዚአብሔር ፍርዱን ይስጠው <i>He has embittered us, may God give him his judgment</i>

Figure 12: Examples of parallel detoxified pairs from AmParaDetox.

Original	老双标狗了 <i>Old double standard dog.</i>
Paraphrase	老是以双重标准看待事物的人了 <i>Someone who always views things with a double standard.</i>
Original	恶心到家了！ <i>Disgusting to the extreme!</i>
Paraphrase	令我不适到家了！ <i>It makes me extremely uncomfortable!</i>
Original	舔狗给她整自信了 <i>The sycophant gave her confidence.</i>
Paraphrase	单方面付出的人给她整自信了 <i>Someone who gives one-sidedly gave her confidence.</i>

Figure 13: Examples of parallel detoxified pairs from ZhParaDetox.

Original	مع انك م**ك بس واففك الرأي <i>Although you are fu**ing bastard, I agree with you</i>
Paraphrase	واففك الرأي <i>I agree with you</i>
Original	فين يا ابن الق**ه السودان مش هيعملوا اي شئ من مخططكم هاهاهاها <i>Where, son of a bit**, Sudan will not do anything of your plan, hahaha</i>
Paraphrase	السودان مش هيعملوا اي شئ من مخططكم هاهاهاهاها <i>Sudan will not do anything of your plan, hahaha</i>
Original	اتفق عليك يا نعال، على كل انا بستاهل <i>D*mn you, you b*stard, anyhow I deserve it.</i>
Paraphrase	على كل انا بستاهل <i>Anyhow I deserve it.</i>

Figure 14: Examples of parallel detoxified pairs from ArParaDetox.