DETOXIFY-IT: An Italian Parallel Dataset for Text Detoxification

Viola De Ruvo[1], Arianna Muti[1], Daryna Dementieva[2], Debora Nozza[1]
[1]Bocconi University, Italy
[2]Technical University of Munich, Germany
viola.deruvo@studbocconi.it
{arianna.muti, debora.nozza}@unibocconi.it
daryna.dementieva@tum.de

Abstract

Toxic language online poses growing challenges for content moderation. Detoxification, which rewrites toxic content into neutral form, offers a promising alternative but remains underexplored beyond English. We present DETOXIFY-IT, the first Italian dataset for this task, featuring toxic comments and their human-written neutral rewrites. Our experiments show that even limited fine-tuning on Italian data leads to notable improvements in content preservation and fluency compared to both multilingual models and LLMs used in zero-shot settings, underlining the need for language-specific resources. This work enables detoxification research in Italian and supports broader efforts toward safer, more inclusive online communication.

***Warning**: this paper contains obfuscated examples some readers may find upsetting and offensive.*[1]

1 Introduction

Online discourse in recent years has become increasingly polluted with toxic language, including hate speech, insults, threats, and other forms of verbal aggression. While automatic systems for content moderation have advanced significantly, most research and resources in this area remain disproportionately focused on English. Detoxification, i.e. the task of rewriting toxic content into a more neutral or non-offensive version, has emerged as a promising alternative or complement to traditional moderation, especially when aiming to preserve user intent while reducing harm. However, the development of robust detoxification systems critically depends on the availability of high-quality, language-specific parallel corpora of toxic and detoxified texts. 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 toxic or neutral, but should not involve deep insults or hate towards individuals or groups of people. While various proactive strategies exist for harmful content moderation—such as countering hate speech (Mathew et al., 2019)—our focus in text detoxification is specifically on mitigating toxic language, particularly targeting less overtly hateful messages.

Original	MA PORCA TR*IA MI CALCOLI CHE C*ZZO
Detox	*Damn it, are you f*cking considering me?* DANNAZIONE MI CALCOLI CHE CAVOLO *Damn it, are you considering me?*
Original	Ma non urtarmi i c*glioni io metto notizie tu mettile a posto e vai a fare in c*lo.
Detox	*Don't bust my b*lls I add the info, you organize it and go f*ck yourself.* Ma non rompermi io metto notizie tu mettile a posto e vai a farti un giro. *Don't annoy me I add the info, you organize it and get out of my face.*

Table 1: Text detoxification parallel pairs examples from our DETOXIFY-IT dataset.

In this paper, we introduce the **first resources and methods for automatic detoxification in Italian**, a language for which there is currently no prior work in this task. We present DETOXIFY-IT, a publicly available parallel corpus containing toxic user-generated comments paired with their manually rewritten, non-toxic versions. By releasing this dataset, we aim to (i) enable the training and evaluation of detoxification systems for Italian, (ii) foster multilingual and cross-lingual research on toxicity mitigation, and (iii) contribute to the broader goal of building safer, more inclusive online environments across languages.

Contributions Our contributions are as follows:

[1]Examples have been obfuscated with a Python package for obfuscating profanities: PrOf by Nozza and Hovy (2023).

Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 267–275

- we release DETOXIFY-IT[2], the first parallel corpus for Italian detoxification at https://github.com/MilaNLProc/detoxify-it;
- we conduct a comprehensive evaluation of state-of-the-art NLP models, including simple baselines, existing multilingual models (both tested via translation and fine-tuned), as well as LLMs.

2 DETOXIFY-IT

This section presents DETOXIFY-IT, a newly created dataset for detoxifying Italian toxic content, consisting of 600 user posts manually rewritten into non-toxic versions. The posts are drawn from three main Italian-language sources: two tweet-based datasets focused on misogyny and homotransphobia detection and a toxicity detection dataset composed of Wikipedia comments. Below, we describe the original datasets and the processing and filtering steps used to build the source material for DETOXIFY-IT. We selected these datasets based on the availability of Italian-language resources and with the goal of varying both the target groups (women and the LGBTQIA+ community) and the domains (Twitter and Wikipedia).

2.1 Twitter Datasets

The two Twitter datasets share the same source platform, as well as similar data collection and annotation procedures. As such, we treat them jointly and apply a unified set of processing and filtering steps.

Both datasets originate from shared tasks at EVALITA, the periodic evaluation campaign for NLP and speech tools in Italian. The misogyny dataset comes from the second edition of the Automatic Misogyny Identification (AMI) shared task at EVALITA 2020 (Fersini et al., 2020). AMI consists of a balanced corpus of 5,000 tweets, collected via keyword searches and by monitoring the accounts of both victims and perpetrators. The hateful posts were subsequently labeled in categories: Stereotype & Objectification, Dominance, Derailing, Sexual Harassment & Threats of Violence, and Discredit. The homotransphobia dataset comes from the first Homotransphobia Detection in Italian (HODI) shared task at EVALITA 2023 (Nozza et al., 2023). HODI contains approximately 5,000 tweets, also collected via keyword searches, and is nearly balanced, with a slight skew toward the negative class.

Each post is annotated as either hate speech or non-hate speech, targeting either women or the LGBTQIA+ community, depending on the dataset. While hatefulness and toxicity are two similar tasks, these labels do not always align: some non-hateful posts may still contain toxic or offensive phrasing, while certain hateful posts are too extreme to be meaningfully detoxified. For instance, some highly toxic content cannot be detoxified due to the lack of a feasible detoxified equivalent, e.g., "mi fa schifo al c*zzo lei e la sua mentalità di merda porca put*ana ma muori" (*en: She makes me sick to the f*ck, her and her sh*tty mentality damn it, you should die*). On the other hand, some posts labeled as non-hateful still exhibit a toxic tone, such as "P*RCA TR*IA RAGA CHE C*ZZO DI LEGGENDA" (*en: DAMN BOY WHAT A F*CKING LEGEND*).

Processing and Filtering In order to obtain a starting dataset to manually detoxify, we needed posts that could be detoxified. Since the available hate speech labels did not reliably indicate detoxifiability, we employed an automatic toxicity classifier to filter the content. Specifically, we used the Perspective API[3], which assigns a score (0 to 1) to various attributes such as Toxicity, Severe Toxicity, Identity Attack, Insult, Profanity, and Threat.

For each post, we retrieved scores across these dimensions and applied threshold-based filtering to discard content that was either too mild or too extreme to meaningfully detoxify. This ensured that only posts appropriate for manual rewriting were retained. See Appendix A for further details. Finally, post length is restricted to 5 to 30 words to maintain readability and contextual clarity.

Following the filtering step, we proceeded to subsample the data for manual detoxification. Since a major part of the original data collection relied on keyword searches, a purely random selection risked overrepresenting certain terms. To mitigate this, we applied stratified sampling based on both the keywords used during data collection and additional high-frequency terms identified in the dataset (see Appendix B).

We then performed stratified sampling to extract 400 posts, balanced across the two target groups: 200 misogynistic posts and 200 targeting the LGBTQIA+ community. Within this sample,

[2]The dataset was used as a part of a test set in TextDetox CLEF 2025 Shared Task (Dementieva et al., 2025b).

[3]https://www.perspectiveapi.com/

we ensured that the relative frequency of each keyword was preserved, maintaining the original distribution. This strategy allowed us to reduce the dataset size while preserving lexical diversity and coverage of different toxic expressions.

2.2 Wikipedia Dataset

The Wikipedia dataset comes from Jigsaw's Multilingual Toxic Comment Classification Challenge[4]. All entries in this dataset are already labeled as toxic, so no further annotation was needed to assess their toxicity level.

Preprocessing and Filtering As with the Twitter datasets, we applied a length filter, retaining only posts between 5 and 30 words to ensure readability, contextual clarity, and to avoid excessively short or long entries. From this filtered set, we randomly selected 200 posts for manual detoxification.

2.3 Annotation Process

We adopted the annotation instructions from Multilingual TextDetox Shared Task (Dementieva et al., 2024b). The main goal of annotation was to ensure that: (i) toxicity is indeed eliminated; (ii) the main content and message of a text are saved as much as possible. Therefore, annotators were instructed to prioritize rephrasing toxic segments, resorting to deletion only when a neutral paraphrase was not feasible.

We manually rewrote 600 toxic texts, balanced across the three source datasets described earlier. The rewriting process was carried out by three native Italian speakers, all with a strong background in NLP and expertise in detecting toxic content. The annotation followed an iterative, collaborative process: one annotator first rewrote the initial 100 toxic texts, after which all three reviewed and discussed the rewrites to resolve disagreements and align with guidelines. This review cycle was repeated after the first 300 and then after all 600 texts.

The final version of the dataset reflects full agreement among the three annotators on each detoxified sentence. Additionally, a fourth expert with experience in NLP detoxification reviewed the entire set, suggesting minor refinements where needed.

[4] https://kaggle.com/competitions/jigsaw-multilingual-toxic-comment-classification

3 Experiments

Given the lack of publicly available detoxification models specifically trained for Italian, we explored several strategies. We started with simple baselines such as toxic word deletion. Then, we evaluated existing models used both in their original form and with translation-based preprocessing, and further fine-tune one of them. Finally, we assessed the zero-shot capabilities of large language models (LLMs) for the detoxification task.

3.1 Baselines

We used the original toxic data as a baseline to assess improvements (**Duplicates**). In the **Deletion** baseline, we took all toxic texts in our dataset and simply removed the toxic words listed in Appendix B. Since our aim is to reduce toxicity while making as few changes to the original sentence as possible, the deletion-based approach represents the most straightforward method for detoxification. It removes explicitly toxic terms without altering the rest of the sentence. While this method does not address implicit toxicity, it provides a useful baseline for comparison with more complex approaches that aim to handle both explicit and implicit toxicity.

3.2 Leveraging Existing Detoxification Models

As mentioned earlier, current detoxification models do not include Italian in their training data. In this section, we evaluate their effectiveness when applied to Italian using three strategies: (i) direct use without modification, (ii) translation-based approaches, and (iii) fine-tuning on our dataset.

Multilingual Transfer (Zero-shot) We first evaluated two multilingual detoxification models (Rykov et al., 2024; Sushko, 2024) introduced in the Multilingual Text Detoxification (TextDetox) 2024 shared task (Dementieva et al., 2024b). These models were trained on parallel corpora in nine languages: English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian, and Amharic, but not Italian. Both models are fine-tuned versions of mt0-XL, differing slightly in their training procedures. We tested both models in a zero-shot setting, using prompts that directly instruct the model to detoxify Italian input while preserving its original meaning. The full prompt templates used in our experiments are provided in Appendix C.

Translation-based Detoxification To further exploit the capabilities of existing detoxification mod-

els trained in other languages, we experimented with a backtranslation pipeline involving English and Spanish. For English, we used the ParaDetox model introduced by Logacheva et al. (2022), which fine-tunes BART on a parallel corpus of toxic and detoxified English texts. For Spanish, we used the same multilingual model evaluated in the previous section (Rykov et al., 2024).

Our pipeline consists of three main steps. First, we translated the toxic texts from the DETOXIFY-IT dataset into English and Spanish using HuggingFace's machine translation models[5]. We opted for these models due to their minimal intervention in tone and meaning, which is critical when dealing with toxic content. More advanced translation systems were avoided, as they often soften or alter the original text, which undermines the detoxification task. Next, we applied the respective detoxification models to the translated texts. Finally, we translated the detoxified outputs back into Italian to complete the process.

Fine-tuning In this experiment, we fine-tuned the model introduced in (Rykov et al., 2024) using our proposed DETOXIFY-IT dataset. For training, we used a total of 300 texts, divided into 240 for training and 60 for validation. These texts are evenly distributed across the three source categories: misogynistic tweets, tweets targeting the LGBTQIA+ community, and toxic Wikipedia comments. The remaining 300 texts are reserved for evaluating the model's performance on the detoxification task.

3.3 LLMs

To investigate whether LLMs can effectively perform detoxification in zero-shot settings, we experimented with two models: Mistral-Nemo-Instruct-2407 and GPT-4o-mini. We used a structured prompt designed to assess both the feasibility and quality of the detoxification process (Appendix C).

3.4 Evaluation

We adopted the multilingual evaluation pipeline from (Dementieva et al., 2024b) for our Italian setup. Following a well-established evaluation framework for text style transfer, we employed metrics to assess three key aspects: (i) the effectiveness of the style transformation from toxic to non-toxic; (ii) the preservation of the original content; and (iii) the fluency and stylistic alignment of the generated output with human references.

Thus, the implementation of the metrics is the following:

Style Transfer Accuracy (STA): Toxicity scores from Perspective API on the generated texts. While Perspective API outputs various labels, we used the results from `toxicity` label which correlates with the toxicity definition we adopted for the annotation.

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

Fluency (ChrF1): ChrF1 score using `sacrebleu` (Post, 2018), measuring closeness to human references.

Joint Score (J): The average of the mean of the above metrics at instance level.

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

where all components are in $[0, 1]$. J is the final metric used for model ranking.

Model	STA	SIM	CHrF1	J
Duplicates	0.421	0.941	0.807	0.323
Deletion	0.740	0.899	0.799	0.534
Backtranslation (EN)	0.795	0.789	0.492	0.318
Backtranslation (ES)	0.852	0.807	0.524	0.370
(Rykov et al., 2024)	0.770	0.900	0.765	**0.542**
(Sushko, 2024)	0.721	0.923	0.776	0.525
Fine tuning	0.624	**0.942**	**0.825**	0.493
Mistral	**0.882**	0.705	0.462	0.306
gpt-4o-mini	0.864	0.854	0.657	0.497

Table 2: Evaluation metrics on DETOXIFY-IT test set. STA for manually detoxified text: 0.677.

4 Results

Table 2 shows the results of our experiments. Using the multilingual model proposed in (Rykov et al., 2024) achieves the highest joint score, suggesting that it offers the best trade-off between toxicity reduction, content preservation, and fluency. Fine-tuning the multilingual model on Italian data significantly boosts SIM and CHrF1, but results in a lower STA, indicating better fluency and meaning preservation at the cost of slightly reduced

[5] `https://huggingface.co/Helsinki-NLP/opus-mt-it-en` for English and `https://huggingface.co/Helsinki-NLP/opus-mt-it-es` for Spanish

Setting	Detoxified
Deletion	ma che , mi sa che ora è il webmaster stesso, ottimo sta solo filando la corda.
Backtranslation (EN)	Ma cosa, penso che sia il webmaster in persona ora, grande è solo girare la corda.
Backtranslation (ES)	Ma cosa, immagino che ora è il webmaster stesso, fantastico sta solo correndo la corda.
(Rykov et al., 2024)	Mi sa che ora è il webmaster stesso, ottimo sta solo filando la corda.
(Sushko, 2024)	Ma che mi sa che ora è il webmaster stesso, ottimo sta solo filando la corda.
Fine-tuning	Ma che cavolo, mi sa che ora è il webmaster stesso, ottimo sta solo filando la corda.
Mistral	Che seccatura, credo che ora sia il webmaster stesso, sta solo prendendo tempo.
gpt-4o-mini	Ma che noia, mi sa che ora è il webmaster stesso, ottimo sta solo tirando avanti la situazione.

Table 3: Original sentence: *Ma che palle, mi sa che ora è il webmaster stesso, ottimo sta solo filando la corda*. Detoxified: *Ma che noia, mi sa che ora è il webmaster stesso, ottimo sta solo filando la corda*. Translation: *What a f*cking pain, looks like it's the webmaster now. Perfect, he's just bailing*.

detoxification strength. Mistral and GPT-4o-mini perform well in reducing toxicity (high STA), but they show weaker fluency or alignment with human references. Backtranslation is the worst approach.

An analysis of the model outputs (Table 3) revealed distinct patterns that helped clarify the results. (Rykov et al., 2024) demonstrates strong performance, although its generated sentences are sometimes ungrammatical. This is partly because toxic elements are removed entirely, which also eliminates the original negative connotation. As a result, the model achieves a higher STA score but lower SIM and CHrF1 scores. In contrast, the fine-tuned model produces outputs that better preserve the negative connotation while detoxifying the toxic content. This leads to slightly higher toxicity scores on average, but they remain comparable to those of manually detoxified sentences (STA = 0.677).

5 Related Work

In the domain of modern NLP for proactive content moderation (Yimam et al., 2024), various strategies have been developed, ranging from fine-grained abusive language classification and text detoxification to counter speech generation. While counter speech with proactive, reasoned arguments is often most effective in addressing severe hate speech, text detoxification techniques are particularly well-suited for moderating content containing profane or offensive language, such as in applications aimed at creating safer online environments for youth (Wachs et al., 2024).

Modern Text Style Transfer (TST) approaches are typically categorized into supervised and unsupervised methods (Jin et al., 2022). Unsupervised models (Dale et al., 2021; Hallinan et al., 2023) have shown strong performance in controllable generation. Recent work has also explored diffusion models for detoxification (Floto et al., 2023; Horvitz et al., 2024) and LLMs for tasks like paraphrasing and detoxification (Zhang et al., 2024). However, models trained on parallel corpora often outperform LLMs, which may hallucinate (Carlson et al., 2018; Rao and Tetreault, 2018; Atwell et al., 2022; Logacheva et al., 2022). Multilingual TST has expanded to a range of languages beyond English. Sentiment transfer has been developed for Indian languages (Mukherjee et al., 2023, 2024), while formality transfer has been extended to Brazilian Portuguese, French, and Italian (Briakou et al., 2021), and to Japanese (Ung, 2023). Detoxification, initially applied to English (Logacheva et al., 2022), has recently been adapted for Russian, Ukrainian, and Spanish (Dementieva et al., 2024a).

While many approaches explored modern LLMs for detoxification on existing and new languages (Toshevska and Gievska, 2025; He et al., 2024; Dementieva et al., 2025a), still there performance is far from being on par with human annotations. Thus, language and cultural specific datasets are highly required for effective proactive text detoxification solutions.

6 Conclusions

This paper presents the first resource for automatic detoxification of Italian texts, introducing DETOXIFY-IT, a manually curated parallel corpus of toxic and detoxified texts. Our evaluation of a variety of approaches demonstrates that fine-tuning a multilingual model with even a small amount of Italian data substantially improves content preservation and fluency, albeit with some trade-off in detoxification strength.

Ethical Considerations

As discussed in prior work, such as ParaDetox (Logacheva et al., 2022), research on toxicity inevitably raises ethical concerns. In particular, one important consideration is the potential misuse of parallel datasets like the proposed one.

While our corpus was created to support the development of systems that reduce harm in online communication, we acknowledge that the parallel structure—containing pairs of toxic and detoxified sentences—could technically be used in the reverse direction, i.e., to "toxify" neutral or non-offensive texts. This raises the risk of generating synthetic toxic content.

However, we emphasize that our dataset was neither designed nor optimized for such reverse use. While we did not conduct a systematic evaluation of reverse detoxification, we align with the observations made by Logacheva et al. (2022), suggesting that applying the process in reverse is unlikely to yield fluent or natural-sounding toxic language. In most cases, the resulting outputs would likely be awkward, semantically inconsistent, or unconvincing as authentic toxic expressions.

7 Limitations

This work comes with a few limitations worth noting. First, the filtering process used to select toxic data from existing hate speech datasets was only partially automated and was finalized through manual review by domain experts. While this helped ensure data quality, it may have introduced some degree of subjectivity.

Second, the annotation process and our own backgrounds as annotators could have influenced the results. However, the involvement of three experts helped reduce individual bias through collaboration and discussion.

Third, we did not experiment with prompt variations when evaluating large language models. Different prompts might produce different outputs, but we believe the overall findings, especially the benefits of language-specific fine-tuning, remain valid.

Acknowlegments

Arianna Muti's and Debora Nozza's research is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101116095, PERSONAE). Arianna Muti and Debora Nozza are members of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis. Daryna Dementieva's work was supported by Alexander Fraser's TUM Heilbronn chair as well as Friedrich Schiedel TUM Think Tank Fellowship.

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.

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.

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.

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, 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. 2025a. 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.

Daryna Dementieva, Daniil Moskovskiy, Nikolay Babakov, Abinew Ali Ayele, Naquee Rizwan, Frolian 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, Vitaly Protasov, Nikolay Babakov, Naquee Rizwan, Ilseyar Alimova, Caroline Brune, Vasily Konovalov, Arianna Muti, Chaya Liebeskind, Marina Litvak, Debora Nozza, Shehryaar Shah Khan, Sotaro Takeshita, Natalia Vanetik, Abinew Ali Ayele, Frolian Schneider, Xintog Wang, Seid Muhie Yimam, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2025b. Overview of the multilingual text detoxification task at pan 2025. In *Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum*. CEUR-WS.org.

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.

Elisabetta Fersini, Debora Nozza, Paolo Rosso, et al. 2020. Ami@ evalita2020: Automatic misogyny identification. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*. CEUR.org.

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

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

Xinlei He, Savvas Zannettou, Yun Shen, and Yang Zhang. 2024. You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 770–787. IEEE.

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.

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.

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.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 369–380.

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 Ondrej Dusek. 2024. Multilingual text style transfer: Datasets & models for indian languages. *CoRR*, abs/2405.20805.

Debora Nozza, Alessandra Teresa Cignarella, Greta Damo, Tommaso Caselli, and Viviana Patti. 2023. HODI at EVALITA 2023: Overview of the Homotransphobia Detection in Italian Task. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)*, Parma, Italy. CEUR.org.

Debora Nozza and Dirk Hovy. 2023. The state of profanity obfuscation in natural language processing scientific publications. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3897–3909, Toronto, Canada. 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, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. 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.

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.

Nikita Sushko. 2024. PAN 2024 multilingual textdetox: Exploring different regimes for synthetic data training for multilingual 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 2892–2900. CEUR-WS.org.

Martina Toshevska and Sonja Gievska. 2025. Llm-based text style transfer: Have we taken a step forward? *IEEE Access*.

Rachel Ung. 2023. *Formality Style Transfer between Japanese and English*. Ph.D. thesis, Waseda University.

Sebastian Wachs, Michelle F Wright, and Manuel Gámez-Guadix. 2024. From hate speech to hateless. the effectiveness of a prevention program on adolescents' online hate speech involvement. *Computers in Human Behavior*, 157:108250.

Seid Muhie Yimam, Daryna Dementieva, Tim Fischer, Daniil Moskovskiy, Naquee Rizwan, Punyajoy Saha, Sarthak Roy, Martin Semmann, Alexander Panchenko, Chris Biemann, et al. 2024. Demarked: A strategy for enhanced abusive speech moderation through counterspeech, detoxification, and message management. *arXiv preprint arXiv:2406.19543*.

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 Source Dataset Processing and Filtering

This section describes the use of Perspective API attributes to filter potential detoxifiable content from the source data: AMI, HODI, and Wikipedia comments. We considered only Toxicity, Identity Attack, and Threat attributes since the other attributes exhibit a high correlation with them.

For Identity Attack and Threat, we set upper limits of 0.5 and 0.2, respectively, to exclude posts that are excessively toxic. These thresholds correspond to the 90th percentile of the score distributions, effectively removing the top 10 % of extreme values. We did not apply a lower bound, as posts with low Identity Attack and Threat scores can still exhibit high overall Toxicity. For instance, the post " Buongiorno... grazie e buon martedì anche a te splendida t*ttona" (*en: Good morning...thank you and happy Tuesday to you too, gorgeous big b*obs girl*) receives low scores for Identity Attack and Threat, yet it is still considered toxic due to its objectifying language.

For the Toxicity attribute, we retain only posts with moderate scores by applying both an upper bound of 0.7 and a lower bound of 0.15. This excludes posts that are either minimally toxic or highly toxic, resulting in a more balanced dataset better suited for detoxification.

Moreover, we discard AMI tweets labeled as sexual harassment, dominance, and discredit due to their high level of hate speech, which could not be detoxified.

B List of Toxic Words

To reduce unnecessary exposure to toxic and hateful language, we do not include the explicit keywords in the main text of this paper. Instead, we provide the full list of keywords used during dataset creation, as well as the set of keywords employed in the deletion-based baseline system, in our GitHub repository at `https://github.com/MilaNLProc/detoxify-it`.

C Experimental Settings

C.1 Fine-tuning settings

The fine-tuning is carried out over 5 epochs with a batch size of 2 and a gradient accumulation step of 4, effectively resulting in a batch size of 8. We use the AdamW optimizer with a learning rate of 5e-5.

C.2 Prompts

Multilingual Transfer Prompt used with (Rykov et al., 2024; Sushko, 2024): "Rewrite the following text to reduce its toxicity while preserving its original meaning:{text}"

LLMs The prompt used with Mistral-Nemo-Instruct-2407 and gpt-4o-mini first asks the model whether the toxic input can be rewritten in a non-toxic way without altering its original meaning. The model is instructed to respond with "yes" if detoxification is possible, or "no" if it is not. If the answer is "yes," it must then generate a detoxified version of the input.

Prompt text : "Can you analyze this example and determine if it can be made less toxic without changing its meaning: TOXIC SAMPLE. Some examples with racist, homophobic, sexist, violent, or personality-targeting content cannot be paraphrased without changing their meaning. Respond with: yes or no. If the answer is 'yes', paraphrase the example to make it less toxic without changing its meaning. Provide the result in the following JSON format: {'response': 'yes', 'paraphrase': PARAPHRASE} or {'response': 'no', 'paraphrase': 'none'}.