To Paraphrase or Not: Efficient Comment Detoxification with Unsupervised Detoxifiability Discrimination

Jing Ke[1], Zheyong Xie[2], Shaosheng Cao[2*], Tong Xu[1*], Enhong Chen[1*]
[1]State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China
[2]Xiaohongshu Inc.
kejing@mail.ustc.edu.cn
{xiezheyong,caoshaosheng}@xiaohongshu.com
{tongxu,cheneh}@ustc.edu.cn

Abstract

Mitigating toxic content is critical for maintaining a healthy social platform, yet existing detoxification systems face significant limitations: *overcorrection* from uniformly processing all toxic comments, and *parallel data scarcity* in paraphrasing model training. To tackle these challenges, we propose DetoxifIability-Aware Detoxification (**DID**), a novel paradigm that adaptively conducts filtering or paraphrasing for each toxic comment based on its **detoxifiability**, namely whether it can be paraphrased into a benign comment in essence. Specifically, DID integrates three core modules: (1) an unsupervised detoxifiability discriminator, (2) a semantic purification module that extracts harmful intents and then performs targeted paraphrasing only on detoxifiable comments, and (3) a feedback-adaptive refinement loop that processes remaining harmful contents only when they are detoxifiable. Experimental results demonstrate that DID significantly outperforms existing approaches on academic data and an industrial platform, establishing a novel and practical modeling paradigm for comment detoxification.

1 Introduction

Commenting on social platforms like Twitter and Xiaohongshu serves as a primary channel for communication and information exchange (Karami et al., 2018). However, malicious comments disseminate harmful content and misinformation, compromising online discourse. Consequently, detoxification of toxic comments has emerged as a critical requirement for social platforms, prompting extensive research into relevant methods (Dale et al., 2021; Hallinan et al., 2023; Lu et al., 2024).

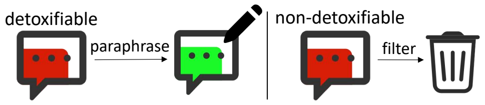

Figure 1: Detoxifiability-Aware Detoxification.

Existing detoxification methods typically filter (Pavlopoulos et al., 2020) or paraphrase (Hallinan et al., 2023) all identified harmful comments regardless of their severity. Such one-size-fits-all approaches induce mismatched intervention intensity, inevitably leading to two failure pathways: (1) deleting all tends to filter slightly offensive instances, resulting in sparse comment sections and diminished user engagement; (2) paraphrasing all harmful content proves ineffective for severe cases (e.g., hate speech and threats), where toxic semantics are intrinsically irreducible to harmless expressions. Although some works have taken preliminary steps to overcome this limitation by introducing **detoxifiability** (Khondaker et al., 2024), namely **whether a malicious comment can be paraphrased into a harmless one**, they suffer from the scarcity of parallel detoxification corpora, namely harmful comments paired with their paraphrased counterparts. The prohibitive cost of human annotation for such data constrains training scale and quality, forming a performance bottleneck and hindering industrial deployment.

To overcome these shortcomings, we take the first step to delve into the DetoxifIability-Aware Detoxification (**DID**) paradigm. DID pioneers the use of the detoxifiability criterion, which assesses whether a comment can be effectively paraphrased into a harmless one to guide subsequent operations, as shown in Figure 1. Built upon the detoxifiability judgments, DID adaptively conducts either paraphrasing or filtering. Concretely, DID consists of three core modules: (1) **an unsupervised detoxifiability discriminator** that assesses whether a harmful comment can be paraphrased effectively; (2) **a semantic purification module** that first extracts harmful intents guided by their toxicity type, then paraphrases only the detoxifiable comments; and (3) **a feedback-adaptive refinement loop** that

[*]Corresponding Authors

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 207–213
March 24-29, 2026 ©2026 Association for Computational Linguistics

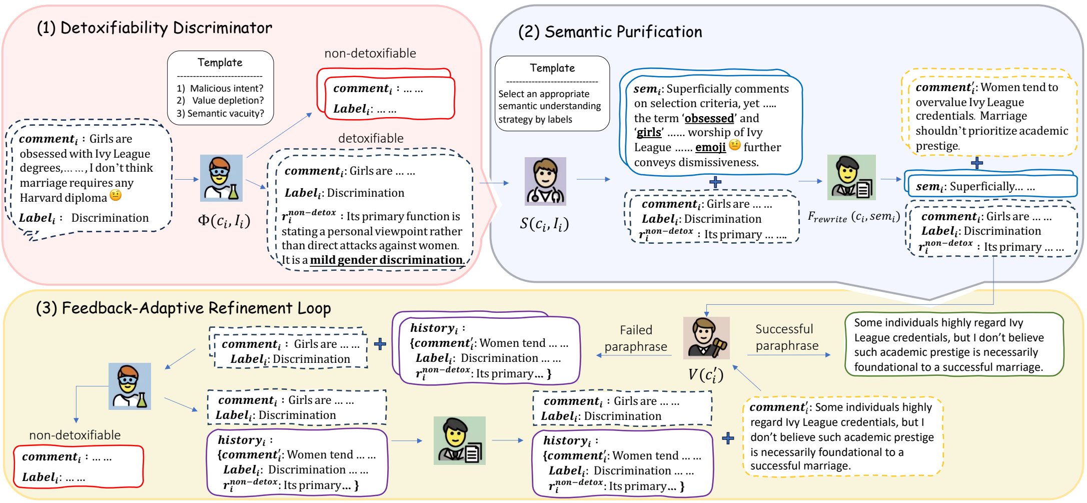
(1) Detoxifiability Discriminator
Template
1) Malicious intent?
2) Value depletion?
3) Semantic vacuity?
non-detoxifiable
comment$_i$: … …
Label$_i$: … …
detoxifiable
comment$_i$: Girls are obsessed with Ivy League degrees,… …, I don't think marriage requires any Harvard diploma
Label$_i$: Discrimination
$\Phi(c_i, l_i)$
comment$_i$: Girls are … …
Label$_i$: Discrimination
$r_i^{non-detox}$: Its primary function is stating a personal viewpoint rather than direct attacks against women. It is a mild gender discrimination.
(2) Semantic Purification
Template
Select an appropriate semantic understanding strategy by labels
$S(c_i, l_i)$
sem$_i$: Superficially comments on selection criteria, yet ….. the term 'obsessed' and 'girls' …… worship of Ivy League …… emoji further conveys dismissiveness.
comment$_i$: Girls are … …
Label$_i$: Discrimination
$r_i^{non-detox}$: Its primary … ….
$F_{rewrite}$ (c_i, sem_i)
comment$'_i$: Women tend to overvalue Ivy League credentials. Marriage shouldn't prioritize academic prestige.
sem$_i$: Superficially… …
comment$_i$: Girls are … …
Label$_i$: Discrimination
$r_i^{non-detox}$: Its primary … …
(3) Feedback-Adaptive Refinement Loop
comment$_i$: Girls are … …
Label$_i$: Discrimination
history$_i$:
{comment$'_i$: Women tend … …
Label$_i$: Discrimination … …
$r_i^{non-detox}$: Its primary… }
Failed paraphrase
$V(c'_i)$
Successful paraphrase
Some individuals highly regard Ivy League credentials, but I don't believe such academic prestige is necessarily foundational to a successful marriage.
non-detoxifiable
comment$_i$: … …
Label$_i$: … …
comment$_i$: Girls are … …
Label$_i$: Discrimination
history$_i$:
{comment$'_i$: Women tend … …
Label$_i$: Discrimination … …
$r_i^{non-detox}$: Its primary… }
comment$_i$: Girls are … …
Label$_i$: Discrimination
history$_i$:
{comment$'_i$: Women tend … …
Label$_i$: Discrimination … …
$r_i^{non-detox}$: Its primary… }
comment$'_i$: Some individuals highly regard Ivy League credentials, but I don't believe such academic prestige is necessarily foundational to a successful marriage.

Figure 2: Illustration of the workflow of our detoxifiability-aware detoxification approach (DID).

verifies the detoxification result and re-invokes the detoxifiability discriminator (1) when necessary.

Experimental results show that DID surpasses existing models by significant margins across various evaluation metrics. Notably, DID outperforms the previous state-of-the-art model by 18.69% in paraphrase success rate and 28.70% in detoxification success rate on real-world data from Xiaohongshu, one of the largest social platforms in China. Moreover, ablation results corroborate the effectiveness of the key modules in DID. We believe that our work will open up a new avenue for studying unsupervised detoxifiability-aware detoxification.

2 Related work

Online comment detoxification (Talat and Hovy, 2016; dos Santos et al., 2018; Schütz et al., 2021) is critical for sustaining the harmony of social media platforms (Sheth et al., 2022). Existing approaches for the task fall into two primary groups: **toxicity detection** and **toxic content paraphrasing**.

Toxicity detection (cjadams et al., 2017; Dubey et al., 2020; Pavlopoulos et al., 2020; Kumar et al., 2024), *a.k.a.* toxicity classification, aims to identify harmful online content for filtration. Besides models making harmful/harmless predictions, some studies (Price et al., 2020; Markov et al., 2023) further predict fine-grained types. However, detection models alone are insufficient for real-world social media moderation, as removing all content deemed harmful risks excessive filtering of mildly toxic samples and engagement decline.

Toxic content paraphrasing models (dos Santos et al., 2018; Pesaranghader et al., 2023; Som et al., 2024) turn offensive texts into non-offensive counterparts while preserving the main semantics. Although advances from GRU (dos Santos et al., 2018) to large language models (Pesaranghader et al., 2023) have significantly improved paraphrasing performance, these approaches share an inherent limitation: they overlook the detoxifiability of harmful comments.While Khondaker et al. (2024) introduced the concept of detoxifiability, their solution necessitates an expensive parallel corpus, limiting its practicality. Moreover, their approach does not explicitly tailor processing based on varying levels of detoxifiability. To our knowledge, we are the first to explore comment detoxifiability identification free from annotated parallel corpora.

3 Methodology

Our **D**etoxif**I**ability-Aware **D**etoxification (**DID**) framework consists of three core modules introduced as follows. Figure 2 illustrates the workflow.

3.1 Preliminaries

We formalize a harmful comment dataset as $\mathcal{D} = \{(c_i, \mathbf{t}_i)\}_{i=1}^{N}$, where c_i denotes the raw text of the i-th comment, and $\mathbf{t}_i \in \{0, 1\}^m$ denotes m-dimensional human-annotated toxicity types, with each dimension indicating a specific harmful type (e.g., dim-0 for hate speech and dim-1 for threats). The pretrained toxicity classifier $\mathcal{V}_{\text{detox}}$ maps each comment c_i to a toxicity probability score $tox_i \in [0, 1]$, where higher values indicate greater toxicity.

3.2 Unsupervised Detoxifiability Discriminator

Prior to detoxification, it is essential to determine comment detoxifiability; bypassing this step risks unnecessary computational overhead and potential harm amplification if detoxification fails. The detoxifiability discriminator $\Phi(\cdot)$ takes a comment c_i and its toxicity type $\mathbf{t}_i$ as inputs and outputs a prediction with explanation. The toxicity type encodes prior knowledge indicating whether the comment is detoxifiable (e.g., harassment comments are typically non-detoxifiable).

The discriminator $\Phi(\cdot)$ predicts whether the paraphrasing is feasible via three criteria:

• *Malicious intent*: Deeply malicious comments cannot be detoxified while preserving semantics.

• *Value depletion*: Comments losing core semantics during toxicity removal are non-detoxifiable.

• *Semantic vacuity*: Content like spam and gibberish should be filtered rather than detoxified.

3.3 Semantic Purification

The semantic purification module consists of two parts—semantic understanding and paraphrasing. It accepts detoxifiable comments, extracts harmful intents and then performs targeted paraphrasing. The details of this process are elaborated as follows.

Semantic Understanding Guided by Toxicity Types Comment detoxification hinges on accurate semantic understanding, which enables both precise localization of toxic elements and faithful preservation of the original meaning. Because many social media comments are nuanced, relying solely on the content makes it quite challenging to identify harmful elements. To tackle this, we leverage the toxicity type $\mathbf{t}_i$ to dynamically select the semantic parsing strategy and infer the intended semantics. For instance, “gender antagonism” comments are more likely to contain deeper implicit meanings through metaphors or double entendres, whereas “harassment” comments typically convey their intent more directly. This type-guided semantic understanding enhances the detection of subtle harmful elements and provides a solid foundation for subsequent semantics-aware paraphrasing.

Semantics-Aware Paraphrasing Leveraging semantic understanding, our paraphrasing module surpasses superficial lexical substitution. It performs sentence-level analysis to identify implicit toxicity and generates the first-round paraphrases c_i'. This process aims to remove toxicity while preserving semantic integrity and fluency.

3.4 Feedback-Adaptive Refinement Loop

To overcome the limits of single-step reasoning, we propose a refinement loop that re-processes failed detoxification cases via re-discrimination and re-paraphrasing.

Formally, we compute the toxicity score of the initial paraphrase c_i' using the toxicity detector $\mathcal{V}_{\text{detox}}(c_i')$. If the toxicity score of c_i' exceeds the preset threshold, it indicates a paraphrase failure.

For failed paraphrases, we leverage historical experiences derived from the past detoxifiability discriminator’s explanation, extracted semantics, and unsuccessful paraphrase attempts. Accumulating diverse failure patterns through historical experience enables effective exploration, preventing models from repeating similar errors. This iterative mechanism progressively incorporates past failure knowledge, significantly enhancing the robustness and efficiency of toxicity mitigation.

4 Experiments

4.1 Evaluation Protocol

We experiment on two datasets, Xiaohongshu Comment and Jigsaw, and their details are provided in the Appendix A. We adopt the following evaluation metrics, grouped by their priority:

• *Primary Task Metrics*: **paraphrase success rate (PSR)**, which is the proportion of comments that are successfully paraphrased among all detoxifiable comments, and **detoxification success rate (DSR)**, namely the proportion of all comments successfully detoxified (i.e., either successfully paraphrased or correctly identified as non-detoxifiable).

• *Foundational Quality Metrics*: **average semantic similarity (ASIM)**, namely the average semantic alignment between a paraphrased comment and the original one, and **average fluency (AFL)**, measuring the grammatical correctness and linguistic naturalness of the paraphrased comments.

The primary task metrics measure the effectiveness of paraphrasing and detoxification, while the foundational quality metrics measure the quality and usability of the paraphrased outputs. Our objective is to maximize PSR and DSR, while maintaining comparable ASIM and AFL scores. Please refer to Appendix B for the implementation details.

4.2 Baselines

We conduct experiments on standard LLM **Llama-3.3-70B-Instruct** (Meta AI, 2024) and reasoning model **DeepSeek-R1** (Guo et al., 2025). For each

Models	Xiaohongshu				Jigsaw			
	PSR↑	DSR↑	ASIM↑	AFL↑	PSR↑	DSR↑	ASIM↑	AFL↑
MaRCo	6.20	6.20	0.6706	3.177	5.40	5.40	0.7537	3.611
DetoxLLM	15.00	15.00	0.6149	4.467	63.10	63.10	0.6625	4.805
Llama-3.3-70B								
Base	47.80	47.80	0.5365	4.887	79.20	79.20	0.6575	4.819
CoT	52.50	52.50	0.5406	4.911	84.80	88.60	0.6575	4.787
DID (Ours)	61.95	77.40	0.5306	4.900	**85.59**	**95.20**	0.6557	4.811
DeepSeek-R1								
Base	45.90	45.90	0.5879	4.723	57.60	57.60	0.6720	4.448
CoT	54.29	63.60	0.5415	4.884	78.21	80.50	0.6667	4.713
DID (Ours)	**72.98**	**92.30**	0.5805	4.723	77.60	87.10	0.6653	4.506

Table 1: The detoxification performance of baselines and our DID approach, where the PSR and DSR metrics are reported in percentage. PSR and DSR are core metrics for assessing detoxification performance.

backbone, we compare our approach against the vanilla baseline (Base) that directly conducts detoxification and Chain-of-Thought (CoT) prompting (Wei et al., 2022) under comparable computation budgets. Additionally, we compare our method with the classic detoxification method MaRCo (Hallinan et al., 2023) and the LLM-based approach DetoxLLM (Khondaker et al., 2024). We give their details in Appendix C.

4.3 Main Results

We present the main results in Table 1 and find that our DID approach outperforms baselines across both datasets. We analyze the failure modes of baselines and the superiority of our DID as follows.

MaRCo mitigates toxicity via character-level replacement. However, due to the inherent brevity and semantic sparsity of social comments, simple text substitution severely compromises fluency and fails to address deep-seated toxicity embedded throughout the text, resulting in limited detoxification efficacy. DetoxLLM requires parallel corpora for training its detoxification model. Consequently, it underperforms significantly on both datasets, especially on the Xiaohongshu Comment dataset, where the recency of the data and domain-specific vocabulary render the fine-tuned model ineffective.

Regarding LLM-based approaches, the CoT method yields significant performance gains over the Base method, achieving the best performance among the baseline models. Our DID model substantially outperforms CoT, raising PSR by 18.69% and 0.79%, and DSR by 28.70% and 6.60% on Xiaohongshu Comment and Jigsaw, respectively. The comparison verifies the effectiveness of our multi-module collaborative framework.

Model Variant	DD	SU	FR	PSR↑	DSR ↑	ASIM↑	AFL↑
Base	✗	✗	✗	47.80	47.80	0.5365	4.887
Ours w/o SU and FR	✓	✗	✗	49.34	69.40	0.5369	4.855
Ours w/o FR	✓	✓	✗	58.29	74.60	0.5278	4.930
Ours	✓	✓	✓	**61.95**	**77.40**	0.5306	4.900

Table 2: Ablation results on the Xiaohongshu Comment dataset using Llama3.3-70B-Instruct.

4.4 Ablation Study

To validate the contribution of the key modules in DID, we conduct ablation experiments and list the results in Table 2. We observe that each module yields significant gains: (1) Adding the detoxifiability discriminator (DD) alone improves PSR by 1.54% and DSR by 21.60%; (2) incorporating semantic understanding (SU) yields further gains of 8.95% in PSR and 5.20% in DSR; (3) the feedback-adaptive refinement loop (FR) further raises PSR by 3.66% and DSR by 2.80%.

5 Conclusion

Online comment detoxification has drawn sustained attention from academia and industry. However, existing approaches either ignore the detoxifiability of toxic comments or rely on labor-intensive parallel corpora for classifier training. In this study, we dissect these inherent drawbacks of existing models and propose **D**etoxif**I**ability-Aware **D**etoxification (**DID**), a novel framework that adaptively conducts filtering or paraphrasing without supervised data. Experiments on academic and industrial benchmarks show that our DID substantially outperforms existing models, achieving state-of-the-art results. As the first detoxifiability-aware method free from costly human annotation, our work offers practical insights for researchers seeking to enhance comment detoxification.

Limitations

To facilitate future research on comment detoxification, we discuss the limitations and possible solutions in this work. (1) Our DID model relies on computing-intensive LLM backbones. We believe that future studies can distill the detoxifiability and paraphrasing capability of LLMs into lightweight models, which will reduce the deployment budget. (2) For processing efficiency and fair comparison with prior works, we only take the comment text as input. We believe that leveraging more context information and collaborative signals will further improve the detoxification performance.

Ethical Considerations

This work studies comment detoxification. As with prior methods, our system may still fail to fully detoxify certain malicious inputs, potentially leaving residual harmful content; nevertheless, our DID approach demonstrates improved effectiveness as shown in the main text. While we do not expect any new risks exposed by our work, we will continue to build on our approach and develop more robust comment detoxification methods in the future.

Acknowledgement

We sincerely thank all the anonymous reviewers for their valuable comments and advice. This work was supported in part by the grants from National Natural Science Foundation of China (No.62222213, U22B2059).

References

cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. 2017. Toxic comment classification challenge. Jigsaw-toxic-comment-classification-challenge. Kaggle.

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.

Daryna Dementieva, Nikolay Babakov, Amit Ronen, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Daniil Moskovskiy, Elisei Stakovskii, and 1 others. 2024. Multilingual and explainable text detoxification with parallel corpora. *arXiv preprint arXiv:2412.11691*.

Cicero 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.

Krishna Dubey, Rahul Nair, Mohd Usman Khan, and Sanober Shaikh. 2020. Toxic comment detection using lstm. In *2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, pages 1–8. IEEE.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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.

Amir Karami, Alicia A Dahl, Gabrielle Turner-McGrievy, Hadi Kharrazi, and George Shaw Jr. 2018. Characterizing diabetes, diet, exercise, and obesity comments on twitter. *International Journal of Information Management*, 38(1):1–6.

Md Tawkat Islam Khondaker, Muhammad Abdul-Mageed, and Laks V. S. Lakshmanan. 2024. DetoxLLM: A framework for detoxification with explanations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19112–19139, Miami, Florida, USA. Association for Computational Linguistics.

Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 865–878.

Junyu Lu, Bo Xu, Xiaokun Zhang, Kaiyuan Liu, Dongyu Zhang, Liang Yang, and Hongfei Lin. 2024. Take its essence, discard its dross! debiasing for toxic language detection via counterfactual causal effect. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15566–15578.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15009–15018.

Jacob Menick, Kevin Lu, Shengjia Zhao, E Wallace, H Ren, H Hu, N Stathas, and F Petroski Such. 2024. Gpt-4o mini: advancing cost-efficient intelligence. *Open AI*.

Meta AI. 2024. Llama 3 model. Technical report, Meta.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305.

Ali Pesaranghader, Nikhil Verma, and Manasa Bharadwaj. 2023. Gpt-detox: An in-context learning-based paraphraser for text detoxification. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1528–1534. IEEE.

Ilan Price, Jordan Gifford-Moore, Jory Flemming, Saul Musker, Maayan Roichman, Guillaume Sylvain, Nithum Thain, Lucas Dixon, and Jeffrey Sorensen. 2020. Six attributes of unhealthy conversations. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 114–124.

Mina Schütz, Christoph Demus, Jonas Pitz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2021. Detox at germeval 2021: Toxic comment classification. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 54–61.

Amit Sheth, Valerie L Shalin, and Ugur Kursuncu. 2022. Defining and detecting toxicity on social media: context and knowledge are key. *Neurocomputing*, 490:312–318.

Anirudh Som, Karan Sikka, Helen Gent, Ajay Divakaran, Andreas Kathol, and Dimitra Vergyri. 2024. Demonstrations are all you need: Advancing offensive content paraphrasing using in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12612–12627, Bangkok, Thailand. Association for Computational Linguistics.

Zeerak Talat and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

A Datasets

We experiment on the following two datasets :

- **Xiaohongshu Comment**: This dataset comprises 8 million real-world comments in Chinese from the social platform Xiaohongshu, annotated with seven toxicity categories: unfriendliness, sexual content, harassment, regional conflict, gender conflict, fandom rivalry, and benign.

- **Jigsaw** (cjadams et al., 2017): This public dataset originates from a toxic comment classification challenge. It contains 300 thousand comments in English labeled as either toxic or benign.

	#Train	#Eval	#Test	ACC↑	F1↑
Xiaohongshu Comment	216k	54	60k	91.32%	90.99%
Jigsaw	5.6k	1.4k	3k	94.29%	94.00%

Table 3: The dataset statistics and the accuracy and F1 scores of the toxicity classifiers.

B Details on Evaluation Criteria

The evaluation of our method is performed on a random sample of 1k toxic comments from each dataset. To evaluate the PSR and DSR, we train toxicity classifiers on each dataset 3 times and report the average results. Besides, we constructed class-balanced subsets for each dataset and trained toxicity classifiers on them to distinguish toxic comments from benign comments. These classifiers were then used to measure the final detoxification effectiveness.

We compute the ASIM using BGE, a widely used embedding model. For Chinese and English texts, we employ the bge-large-zh-v1.5 and bge-large-en-v1.5 models (Xiao et al., 2023), respectively. Additionally, we assess the AFL of the generated samples using GPT-4o mini (Menick et al., 2024), employing a five-level score (higher scores indicate better fluency).

We list the five-level scoring criteria for measuring the fluency of paraphrased text in Table 4.

Score	Definition
5	All sentences are perfectly fluent and natural with no errors.
4	Nearly all sentences are fluent with minor imperfections.
3	Generally fluent with noticeable but non-critical issues.
2	Significant fluency issues affecting comprehension.
1	Severely unnatural with pervasive errors hindering understanding.

Table 4: The scoring criteria of fluency calculation.

C Setups of Baselines

We elaborate on the baseline models as follows:

- **Base** takes four independent detoxification steps to each input to maintain an identical computation budget. The optimal detoxification result is then selected via voting.

- **CoT** follows our method's workflow, guides the model through a step-by-step process: (1) assessing the comment's detoxifiability, (2) extracting semantics, and finally (3) generating rewrite candidates. Finally, select the optimal one from the four candidates via voting.

- **MaRCo** (Mask and Replace with Context) is an unsupervised text detoxification method that combines controllable generation and text rewriting to mitigate toxicity. It uses a non-toxic LM (expert) and a toxic LM (anti-expert) to identify potentially toxic tokens and then replace them.

- **DetoxLLM** is a novel end-to-end framework designed to tackle toxic language across diverse online platforms. It leverages a cross-platform pseudo-parallel corpus generated by ChatGPT for training detoxification models. We load this model to detoxify comments.

D Additional Experimental Results

Although the superiority of our DID approach has been shown on the in-house industrial dataset Xiaohongshu Comment and the public benchmark Jigsaw, widely adopted in prior works (Dale et al., 2021; Dementieva et al., 2024), empirical evidence on more public datasets will further support the conclusion. Therefore, we experiment on the ParaDetox benchmark and present the results in Table 5. We find that our DID substantially outperforms baselines on this benchmark, which shows the generalizability of our approach.

Models	ParaDetox			
	PSR↑	DSR↑	ASIM↑	AFL↑
SmurfCat	56.10%	56.10%	0.8799	3.020
DeepSeek-R1				
CoT	93.84%	94.20%	0.6897	4.837
DID (Ours)	94.30%	96.80%	0.7304	4.820

Table 5: The detoxification performance of baselines and our DID approach, where the core metrics for assessing detoxification performance, PSR and DSR, are reported in percentage.

E Further Discussion

To examine the quality of our detoxification discriminator, we measure the inter-annotator agreement between human judgments and the model's predictions on a randomly sampled subset of Xiaohongshu Comment and record a Cohen's kappa coefficient of 0.7955. The result suggests that our detoxification discriminator achieves near-human accuracy, demonstrating the efficacy of our unsupervised modeling paradigm.

Moreover, to evaluate the fluency assessment capability of GPT-4o Mini, we annotate fluency labels on a randomly sampled subset of Xiaohongshu Comment and compute Cohen's kappa coefficient to measure the agreement between human annotations and model judgments, resulting in a value of 0.6324. The number corroborates the rationality of leveraging GPT-4o Mini for measuring text fluency.