DetoxText: Text Detoxification Using Finetuned Encoder-Decoder Models

M. Serkan Kopuzlu **Kamel Charaf** **Efe Tarhan**
Group 8

Abstract

Toxic language on digital platforms poses a significant threat to constructive communication. However, directly removing such content risks erasing potentially meaningful dialogue and information. In this project, we propose a text detoxification pipeline that rewrites toxic sentences into semantically similar, non-toxic alternatives using fine-tuned transformers. We applied Supervised Fine-Tuning (SFT) on T5 and Qwen models with a composite D-COUNT loss that integrates maximum likelihood, unlikelihood, and toxicity penalties. Additionally, we explore the use of reinforcement learning to further refine detoxification behaviour. Using the ParaDetox dataset, our models outperform existing detoxification baselines across toxicity and semantic similarity. The proposed approaches demonstrate strong balance between content preservation and controllable toxicity reduction, offering adaptable detoxification solutions for platform-specific requirements.

Keywords: Text detoxification, transformer models, SFT, GRPO, toxicity reduction, T5, Qwen.

1. Introduction

The widespread presence of toxic language on digital platforms threatens users' safety, inclusivity, and meaningful discourse. Traditional moderation methods, such as deletion, banning, or keyword filtering, which often lack contextual sensitivity, result in over-censorship or evasion. These limitations have motivated interest in text detoxification, the task of rewriting toxic sentences into neutral forms while preserving semantic content and fluency.

In this project, we address the detoxification task by fine-tuning state-of-the-art language models. We apply our approach to both an encoder-decoder model, T5 [1], and a decoder-only model, Qwen [2]. Both models are trained on the ParaDetox dataset [3], which comprises over 12,000 toxic–neutral sentence pairs. While prior work using maximum likelihood estimation (MLE) has shown promise [4], standard MLE often fails to suppress subtle toxicity or prevent token copying.

To address these issues, we adopt a multi-objective training strategy based on the proposed D-COUNT loss, an extension of the UT loss [5]. While UT loss combines MLE and unlikelihood loss to reduce toxicity in generated text, D-COUNT further incorporates a toxicity classification loss using a frozen Toxic-BERT model. This enhanced formulation enables more effective control over the toxic content while maintaining semantical and structural properties of the sentence. We apply supervised fine-tuning (SFT) with this loss function and empirically tune the loss weights to assess their impact on detoxification quality. Additionally, we integrate reinforcement learning via Group Relative Policy Optimization (GRPO), where we defined our custom reward function based on D-COUNT .

Our contributions are twofold: (1) we validate the effectiveness of D-COUNT in balancing similarity and toxicity reduction through SFT, and (2) we demonstrate the viability of GRPO-based reinforcement learning for enhancing detoxification performance. Our models outperform existing detoxification methods in the literature across both toxicity suppression and semantic similarity metrics.

The rest of this report is organized as follows: Section 2 reviews related work, Section 3 details our methodology, Section 4 presents experiments and results, and Section 5 concludes with future directions.

2. Related Work

Text detoxification has emerged as a focused task within the broader fields of style transfer and harmful language mitigation. Early approaches relied on rule-based systems, which lacked contextual sensitivity and failed to generalize across domains.

With the prime of large pre-trained language models, detoxification systems have become more effective and context-aware. Logacheva et al.[3] introduced the *ParaDetox* dataset, a large-scale parallel corpus of toxic and neutral sentence pairs that enabled the development of supervised detoxification models. Dale et al.[4] benchmarked various transformer-based models, such as T5, BART, and CondBERT, highlighting the trade-offs between fluency, semantic preservation, and toxicity suppression.

EE-559: *Deep Learning*, 2025.

Unlike standard MLE approaches, the original COUNT loss [6] introduces a contrastive objective that jointly leverages likelihood and unlikelihood training to discourage toxic token reproduction.

In parallel, broader efforts in toxic language detection have introduced models like HateBERT [7], which is tailored for abuse detection, and general-purpose transformers like BERT [8] and T5 [1], which have proven effective across generative and classification tasks. Recent advances, such as Qwen [2] and mBART [9], further extend these capabilities to multilingual settings, underscoring the scalability of detoxification frameworks.

3. Method

We propose a two-stage detoxification pipeline that improves over traditional MLE-based approaches by introducing: (1) a novel multi-objective D-COUNT formulation for SFT, and (2) a reinforcement learning stage using GRPO. These contributions address the limitations of toxic content reproduction and enable controllable detoxification without compromising fluency or semantics.

3.1. Dataset and Preprocessing

We use the *ParaDetox* dataset [3], consisting of 12,000+ toxic-neutral sentence pairs collected via crowdsourcing from Reddit, Twitter, and Jigsaw. Sentences are tokenized using the model-specific tokenizers and padded to a uniform maximum length. The dataset is split into 90% training and 10% validation subsets. No further augmentation is performed.

3.2. Supervised Fine-Tuning with D-COUNT

To overcome the shortcomings of conventional MLE objectives-such as toxic token copying, we design a new D-COUNT composed of three components:

MLE Loss ($\mathcal{L}_{\text{MLE}}$)

$$\mathcal{L}_{\text{MLE}} = \mathbb{E}_{(x,y^*)\sim\mathcal{D}} \left[-\log p_\theta(y = y^* \mid x) \right] \tag{1}$$

Maximizes the likelihood of generating a non-toxic paraphrase y^* given a toxic input x, where (x, y^*) pairs are drawn from the parallel detoxification dataset $\mathcal{D}$.

Unlikelihood Loss ($\mathcal{L}_{\text{UL}}$)

$$\mathcal{L}_{\text{UL}} = -\log\left(1 - p_\theta(y = x \mid x)\right) \tag{2}$$

Penalizes the model for generating the toxic input x as the output y, thereby discouraging verbatim copying of toxic phrasing.

Here, x is a toxic source sentence and y^* is its non-toxic reference rewrite. y denotes the sequence generated by the model, which is parameterized by θ. The goal is to generate

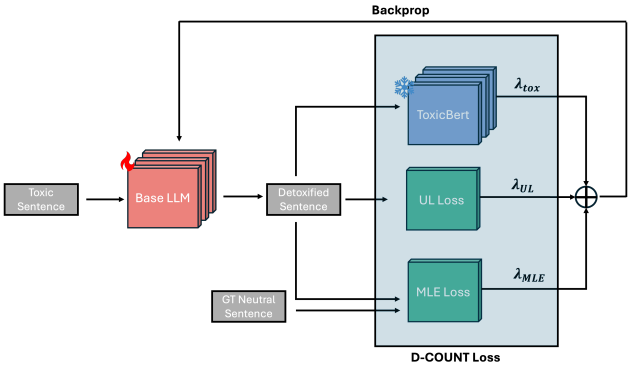

Figure 1: Supervised fine-tuning pipeline with D-COUNT loss objectives.

$y \approx y^*$ while avoiding $y = x$. The dataset $\mathcal{D}$ consists of parallel (x, y^*) pairs used for supervised learning.

Toxicity Loss ($\mathcal{L}_{\text{tox}}$)

$$\mathcal{L}_{\text{tox}} = \text{BERT}_{\text{tox}}(\hat{y}) \tag{3}$$

Uses a frozen Toxic-BERT [7] classifier to penalize high-toxicity generations.

The final objective is a weighted sum:

$$\mathcal{L}_{\text{D-COUNT}} = \lambda_{\text{MLE}}\mathcal{L}_{\text{MLE}} + \lambda_{\text{UL}}\mathcal{L}_{\text{UL}} + \lambda_{\text{tox}}\mathcal{L}_{\text{tox}} \tag{4}$$

We perform a grid search over the coefficients to tune the trade-off between semantic fidelity and detoxification strength. This loss enables adjustable toxicity filtering for real-world applications.

3.3. Reinforcement Learning via GRPO

To complement static supervision, we introduce a GRPO-based training phase [4], where the decoder is modelled as a stochastic policy π_θ. The custom reward function is as followings:

$$\begin{aligned} R(x, \hat{x}) = & \underbrace{\lambda_{\text{sim}} \cdot \cos\left(f_{\text{sim}}(x), f_{\text{sim}}(\hat{x})\right)}_{\text{similarity reward}} \\ & + \underbrace{\lambda_{\text{tox}} \cdot \left(1 - f_{\text{tox}}(\hat{x}) \cdot w_{\text{tox}}\right)}_{\text{toxicity reward}} \\ & - \underbrace{\lambda_{\text{KL}} \cdot \text{KL}\left(\pi_\theta(\hat{x} \mid x) \,\|\, \pi_{\text{ref}}(\hat{x} \mid x)\right)}_{\text{KL penalty}} \end{aligned}$$

The total reward $R(y, x)$ combines similarity, toxicity reduction, and a KL penalty to align generated outputs with the reference model while promoting faithful and non-toxic rewrites.

Baseline and Prior Models				Proposed Methods			
Model	Similarity	Toxicity	C-Score	Model	Similarity	Toxicity	C-Score
Flan-T5$_{\text{BASE}}$	0.68 ± 0.07	0.65 ± 0.13	0.516	**T5$_{\text{D-COUNT, MLE=0.5, UL=0.1, TOX=1.0}}$**	**0.90 ± 0.03**	**0.11 ± 0.07**	**0.893**
mBART$_{\text{TEXTDETOX}}$	0.93 ± 0.03	0.48 ± 0.13	0.727	T5$_{\text{D-COUNT, MLE=0.5, UL=0.5, TOX=0}}$	0.82 ± 0.05	0.63 ± 0.13	0.594
T5$_{\text{PARANMT}}$	0.88 ± 0.03	0.24 ± 0.10	0.824	Qwen3$_{\text{D-COUNT, MLE=0.5, UL=0.5, TOX=0}}$	0.71 ± 0.04	0.10 ± 0.06	0.803
Qwen2.5$_{\text{INSTRUCT}}$	0.55 ± 0.07	0.17 ± 0.10	0.688	Qwen3$_{\text{D-COUNT, MLE=0.5, UL=0.5, TOX=1}}$	0.72 ± 0.05	0.13 ± 0.07	0.794
Qwen3 $_{\text{MLE=0.5, UL=0.5, TOX=0}}$	0.71 ± 0.04	0.10 ± 0.06	0.803	Qwen3$_{\text{GRPO-RL}}$	0.42 ± 0.07	0.26 ± 0.11	0.579

Table 1: Comparison of detoxification models on the ParaDetox dataset. Subscripted model names include key hyperparameters for clarity while preserving space.

4. Validation

As we explained before, our main goal is to strike a balance between similarity and toxicity scores. To enable model comparison, we define a unified metric called the Combined Score (C-score), which equally weights both objectives: C-score $= 0.5 \times (1 - \text{Toxicity}) + 0.5 \times \text{Similarity}$

We began by benchmarking baseline models, including Flan-T5$_{\text{BASE}}$, mBART$_{\text{TEXTDETOX}}$, T5$_{\text{PARANMT}}$, and Qwen2.5$_{\text{INSTRUCT}}$. While mBART$_{\text{TEXTDETOX}}$ achieved the highest similarity (0.93), its toxicity remained high (0.48), resulting in a lower C-score of 0.727. Similarly, T5$_{\text{ParanMT}}$ offered strong similarity (0.88) but still had non-negligible toxicity (0.24), leading to a C-score of 0.824.

In contrast, our fine-tuned T5 models using D-COUNT consistently achieved the best trade-off. The top-performing configuration, T5$_{\text{D-COUNT ,MLE=0.5,UL=0.1,TOX=1.0}}$, reached a similarity of 0.90, toxicity of 0.11, and a C-score of 0.893, outperforming all baseline and prior models. To manage size limitations, we have omitted some rows from the table, but slight variations in the D-COUNT hyperparameters (e.g., higher toxicity weight or MLE scaling) yielded similar performance. We have included the best performing configuration, confirming robustness across different setups.

To further assess the effectiveness of our approach, we compared the performance of our method to that of the UT approach, focusing on toxicity reduction. As indicated in the table, our methods outperform the UT method for the T5 model. This supports the validity of our approach, demonstrating its ability to significantly reduce toxicity while maintaining high semantic similarity, thus proving the robustness and efficacy of our fine-tuning strategy. It is important to note, that in case of Qwen, the scores are relatively similar.

It is important to note that, in the case of Qwen models, the scores are relatively similar rather than showing a clear performance difference, unlike the T5 models. For example, the performance of Qwen3$_{\text{D-COUNT ,MLE=0.5,UL=0.5,TOX=1.0}}$ and Qwen3$_{\text{D-COUNT ,MLE=0.5,UL=0.5,TOX=0}}$ is quite comparable, with only slight variations in toxicity and similarity scores. In contrast, T5 models consistently outperformed Qwen models across all metrics.

Reinforcement learning with GRPO yielded the relatively low toxicity (0.26) but suffered a major drop in similarity (0.42), resulting in the lowest C-score (0.579). These results reinforce our hypothesis that aggressive detoxification via RL may harm semantic fidelity unless carefully balanced.

Overall, D-COUNT models, especially T5, achieved the best trade-off: high semantic preservation, low toxicity, and competitive C-scores (up to 0.893). These findings validate our design choice of multi-objective fine-tuning over traditional MLE or reinforcement learning alone.

5. Conclusion

In this project, we address text detoxification through the lens of controllable fine-tuning. Using T5 and Qwen3 architectures, we introduced the D-COUNT loss, a composite of maximum likelihood, unlikelihood, and toxicity penalties, to guide detoxification while maintaining semantic fidelity.

Our results show that D-COUNT fine-tuning substantially improves the trade-off between toxicity reduction and semantic similarity. The best-performing model, T5$_{\text{D-COUNT ,MLE=0.5,UL=0.1,TOX=1.0}}$, achieved a similarity of 0.90, toxicity of 0.11, and a C-score of 0.893, outperforming all other systems. While Qwen3 variants used D-COUNT, their performance remained similar, with T5 models consistently outperforming them.

Beyond empirical gains, D-COUNT enables flexible detoxification control via hyperparameters, making it suitable for real-world moderation scenarios with varying safety requirements. Future work could explore multilingual extensions, adaptive reward modelling, and alternatives to static toxicity classifiers to improve generalization and fairness.

Ethical Considerations

Toxicity classifiers like Toxic-BERT may reflect societal biases, aggressive detoxification, can distort meaning or tone. While our method improves safety, deployment must prioritize fairness and transparency. Overall, while our work advances technical capabilities for safer online communication, its deployment must be guided by responsible design choices, transparency, and inclusive evaluation practices.

References

[1] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[2] Q. Team, "Qwen3 technical report," 2025.

[3] V. Logacheva, D. Dementieva, S. Ustyantsev, D. Moskovskiy, D. Dale, I. Krotova, N. Semenov, and A. Panchenko, "ParaDetox: Detoxification with parallel data," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Dublin, Ireland), pp. 6804–6818, Association for Computational Linguistics, May 2022.

[4] D. Dale, A. Voronov, D. Dementieva, V. Logacheva, O. Kozlova, N. Semenov, and A. Panchenko, "Text detoxification using large pre-trained neural models," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, (Online and Punta Cana, Dominican Republic), pp. 7979–7996, Association for Computational Linguistics, Nov. 2021.

[5] S. Welleck, I. Kulikov, S. Roller, E. Dinan, K. Cho, and J. Weston, "Neural text generation with unlikelihood training," 2019.

[6] M. M. A. Pour, P. Farinneya, M. Bharadwaj, N. Verma, A. Pesaranghader, and S. Sanner, "COUNT: COntrastive UNlikelihood text style transfer for text detoxification," in *Findings of the Association for Computational Linguistics: EMNLP 2023* (H. Bouamor, J. Pino, and K. Bali, eds.), (Singapore), pp. 8658–8666, Association for Computational Linguistics, Dec. 2023.

[7] T. Caselli, V. Basile, J. Mitrović, and M. Granitzer, "HateBERT: Retraining BERT for abusive language detection in English," in *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, (Online), pp. 17–25, Association for Computational Linguistics, Aug. 2021.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[9] D. Dementieva, D. Moskovskiy, N. Babakov, A. A. Ayele, N. Rizwan, F. Schneider, X. Wang, S. M. Yimam, D. Ustalov, E. Stakovskii, A. Smirnova, A. Elnagar, A. Mukherjee, and A. Panchenko, "Overview of the multilingual text detoxification task at pan 2024," in *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum* (G. Faggioli, N. Ferro, P. Galuščáková, and A. G. S. de Herrera, eds.), CEUR-WS.org, 2024.