International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 04 | April – 2026 DOI:10.55041/ISJEM06274
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Toxic to Positive Comment Rewriting using Supervised Fine-Tuning and Direct Preference Optimization

M. Vardhan
Dept. of Computer Science Engineering
RGUKT Basar
Basar, India B200242

P. Divya
Dept. of Computer Science Engineering RGUKT Basar
Basar, India B200535

G. Krishna Reddy
Dept. of Computer Science Engineering
RGUKT Basar
Basar, India B200596

Abstract—The increased use of harmful, offensive, and disrespectful language online can be attributed to the rapid growth of social media platforms. Although many existing systems focus on detecting and removing harmful content and toxic comments, this approach does not always encourage constructive conversation. Automatically transforming offensive comments into polite and respectful ones, while preserving their original meaning, is a more favorable approach.

Text detoxification is a challenging Natural Language Processing (NLP) task that requires controlled text generation and contextual understanding. In this study, we propose a transformer-based system for rewriting toxic comments that utilizes both Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO).

In the initial phase, parallel toxic and neutral sentence pairs from the ParaDetox dataset are utilized to fine-tune a pre-trained encoder-decoder transformer model (FLAN-T5). This enables the model to learn how to replace harmful expressions with positive alternatives. In the second stage, Direct Preference Optimization is employed to improve alignment with human-generated outputs. By training the model to prioritize more appropriate (neutral) responses over negative (toxic) ones, DPO generates rewrites that are smoother and more semantically accurate.

The system is evaluated using BLEU, ROUGE, sentiment shift, and toxicity reduction metrics. Experimental results show that integrating SFT with DPO improves fluency, politeness, and the preservation of meaning more effectively than using supervised learning alone. This work contributes to the development of intelligent moderation systems that promote safer and more constructive online communication.

Index Terms—Natural Language Processing, Toxic Detoxifica- tion, FLAN-T5, BART, Direct Preference Optimization (DPO), Supervised Fine-Tuning (SFT)

I. Introduction

A. Background and Motivation

The rapid expansion of social media has significantly transformed the way people communicate, share thoughts, and engage in public conversations. Platforms such as Facebook, Instagram, YouTube, Twitter, and other social media sites allow users to connect with a global audience and quickly express their views. While digital platforms have improved communication and the spread of information, they have also led to a noticeable rise in online harmful, aggressive, and offensive language. Aggressive comments, bullying, offensive language, and targeted personal attacks are increasingly com- mon on online platforms.

Harmful content can lead to negative outcomes, including mental distress, cyberbullying, reduced interaction among users, and the spread of negativity in online communities. Most existing moderation systems focus on detecting and removing offensive content. However, simple blocking or removal may not always promote meaningful conversation.

In the field of Natural Language Processing (NLP), the automatic rewriting of harmful comments, also known as text detoxification, has emerged as an important area of research. Detoxification involves generating a new sentence that removes harmful language while keeping the original meaning intact, unlike classification tasks that simply label content as toxic or non-toxic. This task is more challenging because it involves controlled text generation, style transfer, and contextual understanding.

Recent advancements in transformer-based models have significantly improved performance in text generation tasks. Large pre-trained encoder-decoder models, such as T5 and FLAN-T5, have shown strong abilities in understanding con- text and generating smooth, coherent text. Supervised training on its own may not fully align human preferences with model outputs. To enhance alignment and improve output quality, advanced optimization techniques such as Direct Preference Optimization (DPO) are necessary.

B. Challenges in Toxic Comment Rewriting

There are several challenges involved in transforming negative comments into positive alternatives. First, it is essential to maintain the original meaning. The algorithm should remove offensive language while preserving the original meaning of the message. Responses from a poorly designed system can be unhelpful or overly simplistic, leading to a misinterpretation of the intended meaning.

Second, context has a significant impact on toxicity. Some words are not typically disrespectful, but they can convey negative intent depending on the tone and phrasing. Therefore,

instead of relying solely on removing keywords, the model must understand the contextual relationships between words. Third, it is essential to keep using correct grammar and smooth expression. Detoxified outputs should appear natural and human-like. Basic substitution methods can lead to odd or unnatural phrasing. To create writing that is clear and suitable
for the situation, advanced language models are necessary. Another major challenge is preventing excessive sanitization. The model should not completely remove strong beliefs and emotional expressions unless they are harmful. The goal is to promote positive communication instead of imposing restrictions.

C. Related Work and Existing Approaches

Instead of focusing on rewriting, previous research in the area of abusive content processing has primarily centered on the detection of toxic comments. Traditional machine learning methods such as logistic regression and support vector machines (SVM) rely on hand-crafted features, such as bag-of- words models or TF-IDF [2]. These approaches are effective for classification tasks, but since they do not mimic the process of language generation, they are not suitable for generative rewriting tasks.

Transformer models based on encoders, such as BERT and RoBERTa, greatly improved the accuracy of classifying harmful comments due to advances in deep learning [3]. These models use self-attention mechanisms to capture textual context and semantic relationships. Nevertheless, these models are mainly designed for classification and understanding tasks, rather than for generation.

A unified text-to-text framework was established by encoder-decoder transformer designs like T5, BART, and mT5. This paradigm treats tasks like translation, summarization, and paraphrasing as sequence-to-sequence generation prob- lems [3]. Because these models can produce new phrases conditioned on incoming material, they are ideal for text detoxification.

Recent research on alignment and human preference learning has introduced methods such as Direct Preference Optimization (DPO) and Reinforcement Learning from Human Feedback (RLHF) [7]. DPO simplifies preference-based learn- ing by directly adjusting the model to prioritize better outputs over less favorable ones, without requiring an extra reward framework. However, there has been limited research on combining DPO with supervised fine-tuning for the specific purpose of replacing offensive comments.

D. Proposed Work and Contributions

This research introduces a transformer-based framework for transforming harmful comments into favorable ones, utilizing supervised and preference-based optimization methods. The work consists of two primary stages.

First, parallel toxic-neutral sentence pairs from the ParaDetox dataset are utilized to fine-tune aa pre-trained encoder- decoder transformer model (FLAN-T5). By reducing the pre- diction loss at the token level, Supervised Fine-Tuning (SFT) allows the model to learn how to transform harmful language into more positive alternatives.

In the second stage, Direct Preference Optimization (DPO) is employed to move beyond traditional supervised learning methods. The model is trained to prefer a more appropriate (neutral) response instead of a less suitable (toxic) one. This approach enhances the quality of the output in terms of tone and smoothness, while also improving its similarity to text written by humans.

The primary contributions of this work include:

- Developing a FLAN-T5 transformer-based system for rewriting toxic comments.
- Utilizing parallel detoxification data for supervised fine-tuning.
- Incorporating Direct Preference Optimization to improve alignment with human preferences.
- Performing a detailed evaluation using sentiment shift, ROUGE, BLEU, and toxicity reduction metrics.
- Comparing models improved with DPO and those fine-tuned using supervised methods.

E. Organization of the Paper

This paper is organized as follows. Section II presents the dataset and preparation process. Section III explains the methodology and model training. Section IV discusses the encoder-decoder transformer architecture. Section V describes the proposed methodology in detail. Section VI highlights the drawbacks of existing systems and the advantages of the proposed system. Section VII presents the experimental results and evaluation. Section VIII discusses real-world applications. Section IX provides comparative analysis. Section X outlines future scope, and Section XI concludes the paper.

II. Dataset and Preparation

A. Dataset Collection and Filtering

For this experiment, the ParaDetox dataset was utilized, which includes parallel pairs of harmful and neutral sentences [4], [6]. Unlike multilingual classification datasets, this dataset does not need language identification filtering since it has already been curated for text detoxification. To ensure data quality and consistency, several preprocessing methods were employed.

The input sentences were checked for formatting mistakes, unnecessary white space, and incorrectly structured tokens to maintain linguistic accuracy. To avoid bias in training, duplicate entries were removed. Punctuation and emojis were retained because they contribute to the tone and context. This dataset contains aligned sentence pairs, unlike toxic classification datasets, where each harmful comment is paired with a neutral alternative created by a human.

The dataset includes only minimal amounts of personally identifiable information (PII), such as names, phone numbers, or email addresses. In order to maintain ethical standards, any sensitive content identified during the preprocessing stage was removed.

B. Annotation Structure

Rather than relying on binary classification labels, the ParaDetox dataset employs human-written parallel sentence pairs. Each example contains:

- Toxic Comment (Input)
- Neutral / Polite Rewrite (Target Output)

Instead of labeling comments as "toxic" or "non-toxic," the dataset shows the toxic text along with a corrected version. Human annotators created the neutral rewrites to remove harmful or offensive language while preserving the original meaning.

The rewriting process usually focuses on:

- Removing offensive language
- Reducing aggression
- Substituting inappropriate phrases
- Replacing insulting terms
- Enhancing grammatical accuracy

The dataset is well-suited for supervised sequence-to-sequence learning due to its parallel structure.

C. Dataset Characteristics

The ParaDetox dataset has several notable features:

- **Parallel Structure:** Each harmful statement has a matching safe version.
- **Moderate Scale:** The dataset contains approximately 19,744 sentence pairs.
- **Linguistic Diversity:** The sentences include sarcasm, offensive language, and casual phrasing.
- **Sentence Length Variation:** Toxic sentences vary in length from short phrases to long statements, which adds complexity to the generation process.
- **Balanced Tone Transformation:** Instead of imposing restrictions, the dataset focuses on constructive rephrasing.

D. Dataset Splitting

To ensure reliable evaluation, the dataset was divided into three parts: training, validation, and testing sets:

- 80% allocated for training
- 10% used for validation
- 10% reserved for testing

The FLAN-T5 model was fine-tuned using the training set. The validation set was used to adjust hyperparameters like learning rate, batch size, and the number of epochs. BLEU, ROUGE, sentiment change, and toxicity reduction metrics were used on the test set for final evaluation.

III. Methodology and Model Training

A. Overall Framework

The primary aim of this project is to employ transformer-based models to develop an automated system that rewrites toxic content into positive language. The methodology consists of two primary stages:

1) Supervised Fine-Tuning (SFT)
2) Direct Preference Optimization (DPO)

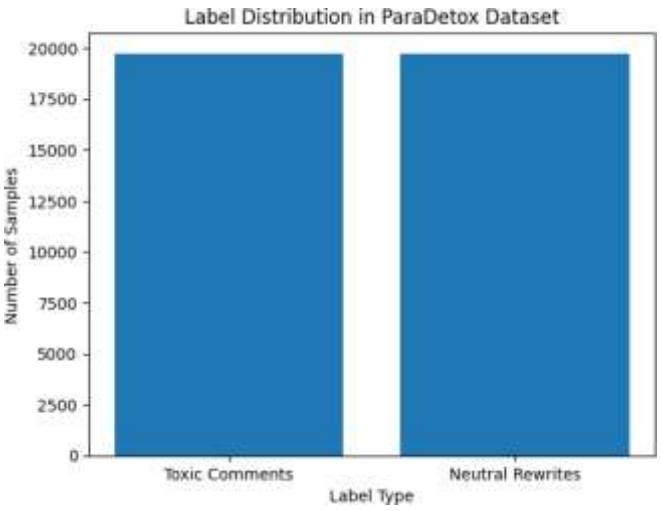

Fig. 1. Label Distribution in ParaDetox Dataset

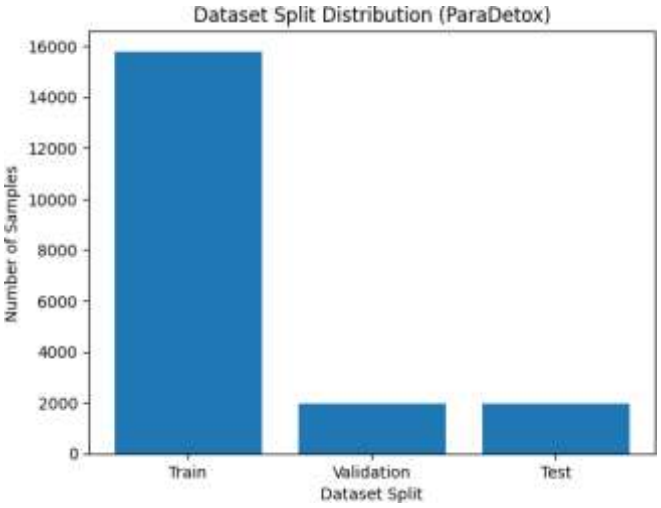

Fig. 2. Dataset Splitting

The entire process involves:

1) Preparing and loading the ParaDetox dataset
2) Tokenizing using the FLAN-T5 tokenizer
3) Supervised fine-tuning of the base model
4) Creating preference pairs for DPO
5) Optimizing based on preference data
6) Evaluation and comparison

B. Data Preprocessing

An essential element in improving model performance is data preprocessing. The following steps were applied:

- Normalization of whitespace
- Removal of duplicate samples
- Preservation of emojis and punctuation
- Use of the FLAN-T5 SentencePiece tokenizer for tokenization
- Conversion of text to token IDs

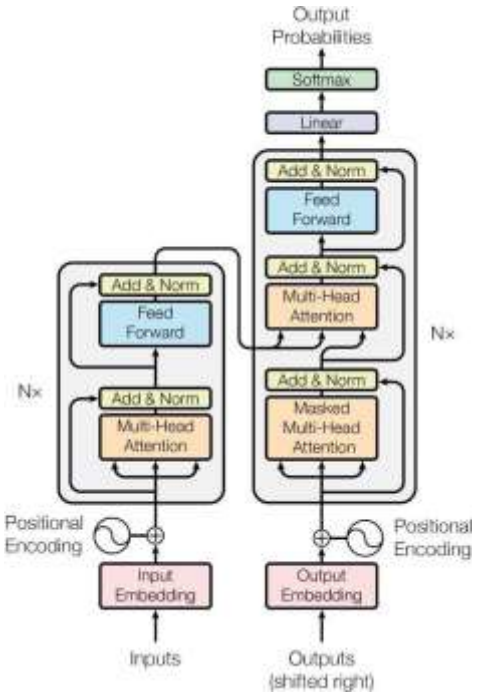

Fig. 3. FLAN-T5 Encoder-Decoder Architecture

Intensive preprocessing techniques, such as stemming or stop-word removal, were not used, unlike traditional machine learning models. Transformer models are designed to automat- ically analyze raw text and obtain contextual representations.

C. Supervised Fine-Tuning (SFT)

In the initial stage, toxic-neutral sentence pairs were utilized to enhance the pre-trained FLAN-T5 encoder-decoder model.
The training process includes:

- The encoder processes the toxic input sentence.
- The decoder generates the neutral version.
- During training, the cross-entropy loss is minimized.

The training setup comprised the following:

- Optimizer: AdamW
- Learning Rate: 5×10^{-5}
- Batch Size: 16
- Number of Epochs: 23
- Loss Function: Cross-Entropy Loss

At this stage, the model is able to directly learn rewriting patterns from parallel examples.

D. Direct Preference Optimization (DPO)

Direct Preference Optimization was employed in the second stage to improve alignment.
Each training example follows this structure:

- Prompt (toxic input)
- Selected response (neutral version)
- Discarded response (original toxic text)

DPO optimizes the model with the aim of making the chosen response more favorable than the discarded one.
The loss function is defined as:

$$L = -\log \sigma (\beta (\log p_\theta(y_c|x) - \log p_\theta(y_r|x)))$$

where:

- y_c represents the selected (neutral) output
- y_r represents the rejected (toxic) output
- β is the preference scaling factor

This stage enhances the fluency, meaning retention, and politeness of the generated outputs.

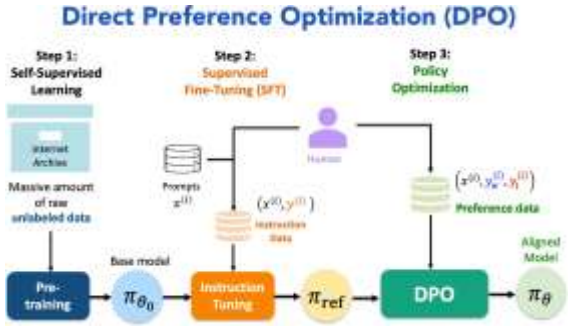

Fig. 4. Direct Preference Optimization (DPO)

IV. Encoder–Decoder Transformer Architecture

A. FLAN-T5 Model Overview

By integrating the self-attention mechanism, transformer-based models have significantly enhanced natural language processing capabilities. Encoder-decoder architectures are well-suited for sequence-to-sequence generation tasks such as text rewriting, unlike encoder-only models which are primarily designed for classification purposes.
The architecture consists of the following components:

- **Encoder:** Generates contextual embeddings by processing the input sentence that contains harmful content.
- **Decoder:** Utilizes the encoder's representations to generate a detoxified output sentence.
- **Self-Attention Mechanism:** Detects the sentence's long-range dependencies.

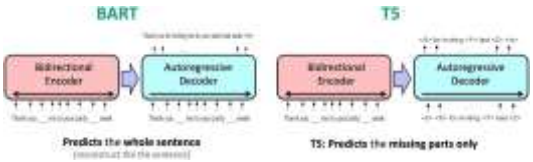

Fig. 5. FLAN-T5 Model

This architecture enables semantic adjustments without altering the original objective, unlike classification methods that only detect toxicity.

V. Proposed Methodology: SFT + DPO

A. Phase 1: Supervised Fine-Tuning (SFT)

During the initial phase, parallel toxic and neutral sentence pairs from the ParaDetox dataset are utilized to enhance the pre-trained FLAN-T5 model. Each training sample includes:

- A harmful input sentence
- A human-generated neutral version

The model is trained using cross-entropy loss, defined as:

$$L_{SFT} = -\sum_{t} \log p_\theta(y_t \mid y_{<t}, x)$$

where:

- x is the harmful input
- y is the neutral target output
- θ denotes the model parameters

Training setup:

- Optimizer: AdamW
- Learning Rate: 5×10^{-5}
- Batch Size: 16
- Epochs: 23

In this phase, the model learns direct rewriting techniques using supervised examples.

B. Phase 2: Direct Preference Optimization (DPO)

While supervised fine-tuning improves the model's ability to rewrite text, it does not explicitly optimize alignment with human preferences. Therefore, Direct Preference Optimization (DPO) is employed.

Each DPO training example includes:

- A prompt x (toxic input)
- A selected response y_c (neutral rewrite)
- A discarded response y_r (toxic output)

The DPO objective is defined as:

$$L_{DPO} = -\log \sigma\left(\beta(\log p_\theta(y_c|x) - \log p_\theta(y_r|x))\right)$$

where:

- y_c represents the preferred (neutral) response
- y_r represents the rejected (toxic) response
- β is a scaling factor

This optimization encourages the model to assign a higher probability to neutral responses and a lower probability to harmful ones. DPO improves fluency, maintains semantic accuracy, and aligns more closely with human preferences.

VI. Drawbacks of Existing Systems

The primary objectives of conventional systems designed to identify harmful comments are classification and blocking. These approaches come with several disadvantages:

- **Restricted Context Comprehension:** Traditional machine learning models depend on TF-IDF features and are unable to grasp the underlying semantic meaning.
- **Hard Filtering Method:** Instead of correcting comments, they are deleted.
- **Absence of Generative Ability:** Classification models are unable to rephrase text.
- **Inadequate Semantic Retention:** Some systems remove content without maintaining its original meaning.
- **No Constructive Moderation:** Removing content does not encourage beneficial interaction.

VII. Advantages of the Proposed System

In comparison to traditional harmful content detection systems, the proposed Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) framework provides significant advantages. The proposed system emphasizes modifying language in a constructive way while preserving its original meaning and intent, unlike classification-only approaches that merely identify or eliminate problematic content.

- **Constructive Rewriting:** This method substitutes harsh or offensive language with polite and respectful terms, rather than removing the words entirely.
- **Semantic Preservation:** The encoder-decoder transformer structure ensures that the original meaning of the message remains intact.
- **Contextual Understanding:** FLAN-T5 employs self-attention mechanisms to capture long-range word relationships within the text.
- **Enhanced Sentiment:** A noticeable improvement in sentiment can be observed in the detoxified outputs.
- **Scalable Deployment:** This method can be integrated using APIs into social media platforms, customer support chatbots, and real-time content moderation systems.
- **Adaptability and Transfer Learning:** The system benefits from transfer learning since FLAN-T5 was pre-trained on large-scale datasets.

VIII. Experimental Results and Evaluation

Traditional measures of classification accuracy are inadequate for evaluating model performance, as toxic comment rewriting involves a text generation task. Instead, fluency, sen- timent improvement, semantic consistency, and the reduction of harmful content are assessed through a range of automated and human-focused evaluation methods.

A. BLEU Score

The BLEU (Bilingual Evaluation Understudy) score measures the similarity between a generated detoxified sentence and a human-written reference sentence by evaluating n-gram overlap. A higher BLEU score indicates that the generated output is more similar to the reference sentence.

B. ROUGE Score

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) emphasizes recall by evaluating the lexical overlap between the reference summary and the generated text. It is particularly useful for assessing whether key terms and phrases from the original source have been preserved. A higher ROUGE score indicates improved information retention and better alignment with the reference text.

C. Sentiment Shift

Sentiment Shift measures the change in sentiment direction from the harmful original text to the cleaned-up version. It assesses the emotional enhancement that occurs after the rewriting process. The sentiment shift is calculated as:

$$\text{SentimentShift} = \text{Sentiment}_{\text{after}} - \text{Sentiment}_{\text{before}}$$

Example:

- Input Sentiment: −0.65 (Negative)
- Output Sentiment: +0.20 (Positive)
- Sentiment Shift: +0.85

A positive sentiment shift suggests a decrease in negativity and an improvement in the overall tone.

D. Toxicity Reduction Rate

Compared to the original input, the Toxicity Reduction Rate measures how effectively the model decreases harmful content in the generated output. A toxicity classifier or an external tool for detecting toxicity is utilized to determine toxicity scores.
The formula for calculating toxicity reduction is:

A.

$$\text{ToxicityReduction} = \frac{\text{Toxicity}_{\text{before}} - \text{Toxicity}_{\text{after}}}{\text{Toxicity}_{\text{before}}}$$

Example:
- Input Toxicity Score: 0.92
- Output Toxicity Score: 0.15
- Toxicity Reduction Rate: 83.7%

A greater reduction rate signifies a more efficient detoxification process.

E. Human Evaluation

Automatic measurements may not fully capture linguistic quality and contextual accuracy on their own. To assess the qualitative aspects of the generated text, human evaluation is conducted.
Human judges use the following standards to evaluate outputs:

- **Fluency:** The naturalness and grammatical correctness of the generated sentence.
- **Meaning Preservation:** Whether the original intent of the harmful input is retained.
- **Politeness:** The degree to which harsh or disrespectful language is reduced.

IX. Real-World Application Scenario

The suggested method for transforming toxic content into positive messages can be incorporated into real-world content moderation and communication platforms. Unlike earlier systems that depend on deletion or blocking, the proposed framework promotes positive communication by converting harmful content into respectful alternatives.

Social Media Integration

In online spaces like discussion forums, comment sections, and community apps, harmful comments frequently result in conflicts and reduced user interaction. The proposed approach can intervene at the time of submission instead of automati- cally deleting offensive comments. Before a harmful comment is posted, the algorithm changes it into a polite and useful version.
Example:
- Input: “You are completely useless.”
- Output: “I believe there is room for improvement in this work.”

This method offers multiple advantages:
- It minimizes online harassment and hostile language.
- It promotes respectful conversation.
- It encourages user engagement rather than suppressing expression.
- It supports healthier online communities.

B. Customer Support Systems

Customer service platforms often encounter messages filled with strong emotions or hostility from unhappy users. Block-
ing these messages directly could lead to increased frustration and harm the overall user experience.
The suggested model is capable of automatically converting customer messages into professional and courteous language prior to their being handled by support agents.

Example:
- Original Message: “Your service is terrible and pathetic.”
- Rewritten Message: “I am dissatisfied with the quality of the service provided.”

This transformation enables:
- Clear expression of concerns.
- Decreased emotional intensity.
- Professional communication between clients and support teams.
- Better management of brand reputation.

These systems can be incorporated into chatbot platforms or business support solutions.

X. Future Scope

The proposed method for rewriting toxic content into positive content employs Direct Preference Optimization (DPO) and Supervised Fine-Tuning (SFT) to achieve promising re- sults. However, several aspects of the framework, including scalability, applicability, and robustness, can still be enhanced further.

A. Advanced Large Language Models

Future research could explore the use of more sophisticated and larger transformer architectures, such as LLaMA, GPT-style models, or multilingual versions of T5. These models may generate outputs that are smoother, have more detailed semantic representations, and demonstrate improved contex- tual understanding.

B. Multilingual Detoxification

Currently, the system primarily centers on English text. Expanding the platform to support multilingual detoxification, including Indian languages and code-mixed content, would greatly enhance its practical value. Multilingual fine-tuning could enable cross-lingual generalization.

C. Fine-Grained Style Control

Future studies could incorporate mechanisms for controllable text generation. Instead of simply generating a neutral rewrite, the system could offer customizable tone settings such as:

- Moderately polite
- Highly professional
- Empathetic tone
- Constructive feedback style

This would allow users to choose from various rewriting styles tailored to their needs.

D. Reinforcement Learning and Preference Scaling

While DPO improves alignment, further research could explore:

- Approaches for adaptive preference scaling
- Reinforcement learning that incorporates human feedback during the learning process
- Continuous learning through user corrections

This can further enhance alignment with evolving standards of human communication.

E. Explainable and Trustworthy AI

Transparency can be enhanced through the use of explainability methods such as attribution analysis and attention visualization. When users and moderators can understand the reasoning behind changes to specific terms, it increases their confidence in automated moderation systems.

F. Robustness and Bias Evaluation

Future research should examine possible biases in rewriting methods and assess the model's ability to withstand adversarial inputs. To implement AI responsibly, it is crucial to guaran- tee fairness and avoid unintentional misinterpretation of user intentions.

G. Human-Centered Evaluation Expansion

Expanding large-scale human evaluation studies to include diverse demographics can provide deeper insight into percep- tions of politeness, cultural sensitivity, and real-world usability. The practical impact and research contribution of the pro- posed SFT + DPO framework will significantly grow if it is extended to handle multilingual scenarios, controllable generation, scalable implementation, and ethical AI concerns.

XI. CONCLUSION

This research introduces a transformer-based approach for transforming toxic comments into positive ones, employing Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO).

The proposed approach focuses on modifying language in a constructive way while preserving the original meaning, unlike traditional methods for identifying toxic comments, which depend on classification and removal.

First, parallel toxic-neutral sentence pairs from the ParaDetox dataset were utilized to fine-tune a pre-trained encoder-decoder model (FLAN-T5). Through supervised fine- tuning, the model learned to rewrite text by removing objec- tionable expressions while maintaining grammatical correct- ness and preserving the original meaning in context.

In the second stage, Direct Preference Optimization was employed to improve alignment with human-generated outputs.

The proposed solution offers a practical and adaptable alternative to traditional moderation methods. By replacing inap- propriate language with respectful and meaningful alternatives, it promotes positive communication instead of restricting user content. This research supports the development of intelligent moderation algorithms that align with human values, aiming to improve online discussions on social media and other communication platforms.

This technology could see future advancements such as multilingual detoxification, customizable tone generation, real-time deployment optimization, and improved alignment through advanced preference-based learning methods.

ACKNOWLEDGMENT

The authors would like to sincerely thank **Mr. Sujoy Sarkar**, Department of Computer Science Engineering, RGUKT Basar, for his valuable guidance, continuous support, and encouragement throughout the completion of this project and paper.

REFERENCES

[1] A. Pesaranghader, N. Verma, and M. Bharadwaj, "GPT-DETOX: An In-Context Learning-Based Paraphraser for Text Detoxification," arXiv preprint arXiv:2404.03052, 2024.

[2] S. Mukherjee, A. Bansal, A. K. Ojha, J. P. McCrae, and O. Dus˘ek, "Text Detoxification as Style Transfer in English and Hindi," arXiv preprint arXiv:2402.07767, 2024.

[3] L. Laugier, J. Pavlopoulos, J. Sorensen, and L. Dixon, "Civil Rephrases of Toxic Texts with Self-Supervised Transformers," in Proceedings of the European Chapter of the Association for Computational Linguistics (EACL), 2021.

[4] N. Vanetik, "A Dataset for Offensive Language Detoxification," in Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP), 2025.

[5] X. Yi, L. Wang, X. Wang, and L. He, "Fine-Grained Detoxification via Instance-Level Prefixes for Large Language Models," arXiv preprint arXiv:2402.15202, 2024.

[6] D. Dementieva, N. Babakov, and A. Panchenko, "MultiParaDetox: Extending Text Detoxification with Parallel Data to New Languages," arXiv preprint arXiv:2404.02037, 2024.

[7] X. Xie, T. Li, and Q. Zhu, "Learning from Response not Prefer- ence: A Stackelberg Approach for LLM Detoxification," arXiv preprint arXiv:2410.20298, 2024.