Detoxifying Online Discourse: A Guided Response Generation Approach for Reducing Toxicity in User-Generated Text

Ritwik Bose[α,β] and **Ian Perera**[α] and **Bonnie J. Dorr**[φ]
α: Florida Institute for Human and Machine Cognition, β: Knox College,
ϕ: University of Florida
rbose@ihmc.org iperera@ihmc.org bonniejdorr@ufl.edu

Abstract

The expression of opinions, stances, and moral foundations on social media often coincide with toxic, divisive, or inflammatory language that can make constructive discourse across communities difficult. Natural language generation methods could provide a means to reframe or reword such expressions in a way that fosters more civil discourse, yet current Large Language Model (LLM) methods tend towards language that is too generic or formal to seem authentic for social media discussions. We present preliminary work on training LLMs to maintain authenticity while presenting a community's ideas and values in a constructive, non-toxic manner.

1 Introduction

Certain topics, beliefs, and views can lead to polarized and inflammatory discourse, often with little headway made in conveying these views to an opposing side in a constructive manner. As such topics are often strongly tied to emotional experiences, beliefs, and identity, the expression of such views likewise carry heavy emotional weight which can hinder constructive discourse. We thus explore the potential for natural language generation through Large Language Models (LLMs) to mediate such discourse by expressing polarizing views and opinions in a manner with reduced toxicity and with a style more closely mimicking that of the target audience.

Consider the following human authored Reddit post: *There's no shame in abortion. Only repressed and self loathing people like you, feel that way. I've had two and don't regret anything. Your shame tactics don't work with me, cupcake.* The text is associated with computed levels of *interaction measures*: toxicity, sentiment, and attitude. From this information, language generation can be applied to the goal of encouraging constructive discussion through an application of large language models to attempt to retain the underlying meaning while removing inflammatory or divisive language.

Specifically, a 3-layer system couples an encoder-decoder model and a language generation model, where the encoder-decoder identifies content parameters from a post. These, taken together with the original subreddit and the interaction measures to fill in blanks to create a synthetic post: *I have had two abortions and stand by my choices. Your attempts to shame me are ineffective, my friend. It is not productive to sham people for their own decisions. You should be ashamed of your own choice. Abortion is a personal choice that should not carry any negative stigma. Everyone is entitled to their own beliefs, but shaming others for their own decisions is not productive. Personally, I have had two abortions and stand by my choices. If you want to know more about me, please feel free to reach out to me.*

Many current LLM methods, such as ChatGPT[1], tend towards a tone of summarization while also generating text that is prosaic without a prompt specifically asking for a specific style. This tone is likely to be seen as that of a moderator or automated system, and we believe would less effective in improving discussions in communities.

Our approach of learning community style and opinions while mediating toxic expression supports our ultimate goal of an *echo-chamber burster*, where we generate posts resembling a subreddit's style but presenting an opposing view point. The assumption here is that by constructively expressing an opposing view, the community may be more open to listening to the opposing view. We anticipate this system could be used for social media platforms, moderators, or widgets to suggest alternative phrasings of posts that maintain the underlying views of an individual while moving towards more constructive community engagement. To this

[1] https://chat.openai.com

Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023), pages 9–14

end, we make our system available for public use[2].

2 Prior Work

One can partially frame the problem of reducing toxicity while retaining semantic content as a style transfer problem. Tokpo and Calders (2022) perform style transfer to mitigate bias training on non-parallel texts by mapping from the latent space of biased text to non-biased text. Reif et al. (2022) explores the efficacy of zero-shot, one-shot, and few-shot prompts for style transfer. Adversarial approaches (Chawla and Yang, 2020; Fu et al., 2018) have also shown strong results when applied to parallel data. However, these methods tend to prioritize fluency and sometimes formality, which could be seen as inauthentic in social media discussions.

Several recent studies have explored various aspects of large language models (LLMs) and their applications to similar problems. Sadasivan et al. (2023) investigated the automated detection of LLM-generated text and found that such detection can be obfuscated by paraphrasing the LLM output using a lighter T5 model. Bhaskar et al. (2022) demonstrated that GPT-3 provides an inherent level of "factualness" and "genericity" when summarizing collections of reviews.

Moreover, the adaptability of ChatGPT to different cultural contexts was assessed by Cao et al. (2023). They determined that while ChatGPT is capable of adapting to an American cultural context, it encounters difficulties with other cultural contexts. This limitation poses challenges when attempting to adapt the model to individual communities, especially those that are orthogonal to each other, such as `r/prochoice` and `r/prolife`.

Wei et al. (2022) examined the capabilities of LLMs in zero-shot learning scenarios. While LLMs exhibit impressive zero-shot learning abilities, their findings suggest that fine-tuned models, when combined with tailored prompts, are more effective at generating the desired outputs. This insight is particularly relevant to our detoxification task, where we employed a fine-tuned model with crafted prompts to guide response generation.

We observe that major large conversational models (eg ChatGPT, Bard, Claude) are both closed sourced and constantly undergoing improvement, rendering prompts unstable. Our approach does not rely on sophisticated prompt engineering and can work on controversial domains with profanity.

[2]https://github.com/infiniterik/detoxify

3 Data Collection

Data was gathered from Reddit using a data collection and enrichment framework that combined multiple collection methods to ensure coverage over different timescales and moderator activity.

We used a combination of the PushShift API [3] and the Python Reddit API Wrapper (PRAW) [4] for collecting posts and comments from Reddit communities. While PushShift can efficiently return cached comments and posts, it does not provide updated upvote/downvote data – we thus obtain revised scores using PRAW. Additionally, we collect data repeatedly each day to determine which comments and posts have been removed by moderators using the method described in Chandrasekharan and Gilbert (2019).

We used data gathered from `r/prochoice` and `r/prolife` over the span of a year in order to gather contrasting and opposing viewpoints. Including posts and comments which were deleted, we gathered 116,293 items from `r/prochoice`.

Each comment and post is enriched with off-the-shelf tools for classifying text based on emotion[5], sentiment[6], and toxicity (Hanu and Unitary team, 2020). A summary of each post was also included using a fine-tuned version of `flan-t5-xxl`[7].

For the final experiments a randomly selected training (70%) and test (20%) split was constructed. In order to avoid cross-contamination between training and test data, child posts were only allowed in the training or test set if the parent post was also in the same set. The training set contained 65,292 posts labelled `low toxicity` and 11,936 posts labelled as `high toxicity` while the test set contained 18,631 `low toxicity` posts and 3,435 `high toxicity` posts.

4 Methods

We conducted a detoxification task aimed at rephrasing posts with high toxicity scores to reduce their toxicity while preserving the author's original intent. To ensure that the appropriate context was provided for generating the target post, we incorporated summaries of parent post and the target post as a part of the prompt for our model. We

[3]https://files.pushshift.io/reddit/

[4]https://github.com/praw-dev/praw

[5]https://hf.co/bhadresh-savani/distilbert-base-uncased-emotion

[6]https://hf.co/nlptown/bert-base-multilingual-uncased-sentiment

[7]https://hf.co/jordiclive/flan-t5-11b-summarizer-filtered

Experiment	Prompt Text
parent-child with summaries (PCS)	Post summary: `?parent_summary`. A post: `?parent_post`. Reply summary: `child_summary` A reply:
parent-child with toxicity and summaries (PCTS)	Post summary: `?parent_summary`. A `parent_toxicity` post: `?parent_post`. Reply summary: `?child_summary` A `?child_toxicity` reply:

Table 1: We constructed prompt-completion tasks for to fine-tune a `T5-Large` model over. The `?parent_toxicity` and `?child_toxicity` levels were identified by thresholding the toxicity scores for each post. Summaries were automatically identified using a `T5`-based model. The target text for each prompt was the content of the `?child_post`.

framed our detoxification task as a guided response generation task and employed a fine-tuned model to replicate the tone of the designated subreddit. In this approach, we utilized prompt-completion pairs created using templates, as detailed in Table 1. We refer to a post responding to a previous post as the child post, while the post being responded to is considered the parent post. To avoid cross-contamination between test and training sets, we discarded any posts in which parent and child pairs were not present together in the same training or test split.

4.1 Enrichment Encoding

In order to translate numerical enrichment data into text, we picked a threshold α and labelled all posts with a value less than α toxicity score as `low toxicity` and all posts with greater than α toxicity score as `high toxicity`. We found $\alpha = 0.5$ to be appropriately discriminative as 81% of posts had a toxicity score of less than 0.33 and 12% of posts had a toxicity score greater than 0.66. The thresholded values then feed into prompts to train the models, as described below.

For the ChatGPT implementation, we created a comparable test set of detoxified posts using the prompt "*Rephrase the following Reddit post to be less toxic:* `?child_post`".

4.2 Template Construction

We fine-tuned `T5-large`[8] on our specific completion tasks which are framed as prompts for a comment responding to a given post. We constructed several different completion tasks using a combination of parent and child data.

- **PC**: The parent post is provided with the goal of producing the child post.
- **PCS**: The parent post and summaries of the parent and child post are provided.
- **PCTS**: Same as PCS with toxicity label for each of parent and child.
- **PCTS+ChatGPT**: Same as PCTS using ChatGPT rephrasing instead of summaries

Table 2 shows an example of a given parent post, the original response text, and the outputs from different system configurations.

4.3 Fine-Tuning

We used a modification of the `SimpleT5` library[9] to enable multi-gpu training. For each template, we trained a different model for 5 epochs. We tested three different configurations of LLMs with and without toxicity training data, and summarization provided either by the fine-tuned `flan-t5-xxl` model or ChatGPT.

5 Evaluation and Results

5.1 Automatic Evaluation

Rephrased responses were evaluated for semantic similarity, measured using BLEURT (Sellam et al., 2020). We use the original human response as the reference in order to measure the deviance in meaning due to rephrasing. The PC and PTS tasks do not contain direct content guidance for the target post, as they consist only of the parent post and parent post with summary and toxicity respectively. Results from automatic evaluation are shown in Table 3.

5.2 Human Evaluation

To assess the AI-generated responses, annotators were instructed to rate the outputs of the PCS, PCTS, and PCTS+ChatGPT systems on a scale of 1-5 across four categories: **fluency**, **toxic-constructive scale**, **authenticity/humanness**, and **relevance**. Fluency refers to the comprehensibility and cohesiveness of the generated text, while the toxic-constructive scale measures the degree

[8] https://hf.co/t5-large

[9] https://github.com/Shivanandroy/simpleT5

System	Output
Parent	*People are not incubators!*
Original Child	*Indeed. These anti-abortion bills are horribly sexist.*
PC	*[deleted]*
PCS	*These anti-abortion bills are gonna make us all sick.*
High Tox PCT	*I'm not a person, but I am an incubator.*
Low Tox PCT	*[deleted]*
High Tox PCTS	*These anti-abortion bills are f***ing insane.*
Low Tox PCTS	*These anti-abortion bills are causing women to die from unsafe abortions.*
ChatGPT	*It's true that these bills aimed at restricting abortion are discriminatory towards women.*
PCTS+ChatGPT	*It's true that these bills aimed at restricting abortion are discriminatory towards women. But I don't think it's fair to say that they haven't done any research on the actual effects of those laws, and how they could be used as a way to make them more difficult to get an abortion.*

Table 2: Example System outputs for a parent-child post pair.

Experiment	High	Low
chatGPT+PCTS	0.357	0.366
PCTS	0.433	0.438
ChatGPT	0.432	
PTS	0.201	
PCS	0.456	
PC	0.196	

Table 3: Average BLEURT scores for different models. The PCTS models are able to select either high or low toxicity. We see that while the selected toxicity does not significantly impact the total BLEURT scores.

	Fluency	Tox/Con	Auth	Rel
PCS	**4.06**	2.57	**3.79**	3.14
PCTS	3.96	2.54	3.68	3.25
PCTS+ ChatGPT	3.92	**3.32**	3.39	**3.71**

Table 4: The results of annotations. 5 annotators rated output posts for fluency, toxicity/constructiveness, authenticity, and relevance.

Fluency	Tox/Con	Auth	Rel
0.46	0.44	0.35	0.44

Table 5: Inter-annotator agreement calculated as the Krippendorff Alpha using interval metric.

to which a post could either contribute to hostility or foster constructive dialogue. Authenticity/humanness rates how natural and human-like the AI-generated response appears, and relevance evaluates the appropriateness of the response in relation to the parent post. Annotators rated 300 instances divided across the three configurations, with each instance rated by two annotators. Results from the evaluation are shown in Table 4. Inter-annotator agreement was calculated using Krippendorff's Alpha with the interval metric and is shown in Table 5.

6 Discussion

Our results show that LLMs fine-tuned on communities can lead to more authentic generated text, but can learn toxic response patterns without measures to reduce such toxicity. Additionally, existing toxicity rating libraries can provide a helpful signal to reduce toxicity, albeit with limitations. While the guidance from ChatGPT improves constructiveness and relevance, we see that authenticity is maximized with the fewest additions to the underlying LLM. In the automated evaluation, we observe that PCTS, which includes information about the parent post (text, toxicity, and summary) performs on par with ChatGPT-only detoxification; but using ChatGPT to produce summaries for PCTS does not accumulate the benefits. Additionally, we find that reducing toxicity does not strongly affect the BLEURT score, which is expected, but also demonstrates that BLEURT is invariant to differences of constructiveness and sentiment of text content.

7 Future Work

While augmenting and guiding response generation with summaries was essential in binding the

output to the original text, summarization alone is insufficient in maintaining deeper meaning. Certain opinions or statements can lead to uncivil discourse in a community even absent profanity or negative language directed towards individuals. We aim to address this using additional guidance in the form of stance predicates (Mather et al., 2022) to improve the faithfulness of detoxified posts to their original text.

8 Limitations

8.1 Toxicity measure

The automated suite of enrichments that we used does produce erroneous output, often conflating profanity with toxicity. Additionally, the emotionally and politically charged nature of our dataset lends itself to potentially subjective measures of toxicity. Ultimately, our idea of a toxic post would be a post which violates community standards and leads to discord in the community. Additionally, while we can fail to reduce toxicity, a deeper study should be done to determine whether we run the risk of unwittingly increasing toxicity under certain situations. Additional research regarding measures of community health is forthcoming and will address the appropriateness of the toxicity measures.

8.2 Resources and language availability

The `T5-large` models were trained on A6000 GPUs. Further optimization to reduce resource requirements is possible. Models were limited to 512 tokens, meaning longer posts may be poorly rephrased. Performance of the system in non-English languages depend on availability and performance of the enrichment and summarization models in those languages.

8.3 Evaluation and Annotations

Stronger claims about the unique characteristics of our approaches would require more robust evaluation methods and additional domains. The Likert scale approach for rating system output suffers from such drawbacks – such as differences in interpretation of the scales and a tendency to choose middle rating in uncertain cases. Ranking system outputs according to these scales would address some of these limitations, but would increase annotator time commitment and lose the magnitude of quality differences between models. Other methods, such as system-level probabilistic assessment (SPA), could potentially provide a superior evaluation (Ethayarajh and Jurafsky, 2022).

Ethics Statement

Given the sensitive nature of our target domain and problem space, this technology has several potential ethical implications and considerations. First, creating detoxified text effectively raises the possibility of creating extra toxified text as well – our system can produce more toxic text which could be used to produce a falsified perspective on a community. Increasing perceived authenticity also increases potential for misuse, as many current methods for detecting AI-generated text may be thwarted by these methods. Additionally, our measure of toxicity is currently limited by an external system – these libraries are often slow to update to current events, memes, and new language that can be used in a toxic manner. However, the ability to tune an LLM for a particular community could conversely provide a means to learn such patterns in language use.

Overall, we believe the potential for this work to aid in generating constructive discussions outweighs the potential harms from its misuse.

Posts from Reddit were automatically deidentified, and work was performed with approval from our institution's IRB.

Acknowledgements

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR00112290022. The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

Adithya Bhaskar, Alexander R. Fabbri, and Greg Durrett. 2022. Zero-shot opinion summarization with gpt-3.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*.

Eshwar Chandrasekharan and Eric Gilbert. 2019. Hybrid approaches to detect comments violating macro norms on reddit.

Kunal Chawla and Diyi Yang. 2020. Semi-supervised formality style transfer using language model discriminator and mutual information maximization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2340–2354, Online. Association for Computational Linguistics.

Kawin Ethayarajh and Dan Jurafsky. 2022. The authenticity gap in human evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6070, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style Transfer in Text: Exploration and Evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Brodie Mather, Bonnie Dorr, Adam Dalton, William de Beaumont, Owen Rambow, and Sonja Schmer-Galunder. 2022. From stance to concern: Adaptation of propositional analysis to new tasks and domains. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3354–3367, Dublin, Ireland. Association for Computational Linguistics.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A Recipe for Arbitrary Text Style Transfer with Large Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Ewoenam Kwaku Tokpo and Toon Calders. 2022. Text Style Transfer for Bias Mitigation using Masked Language Modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned Language Models Are Zero-Shot Learners.