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

Ritwik Bose[α,β], **Ian Perera**[α], **Bonnie J. Dorr**[ϕ]

[α]Institute For Human and Machine Cognition

[β]Knox College, [ϕ] University of Florida

{rbose, iperera}@ihmc.org, bonniejdorr@ufl.edu

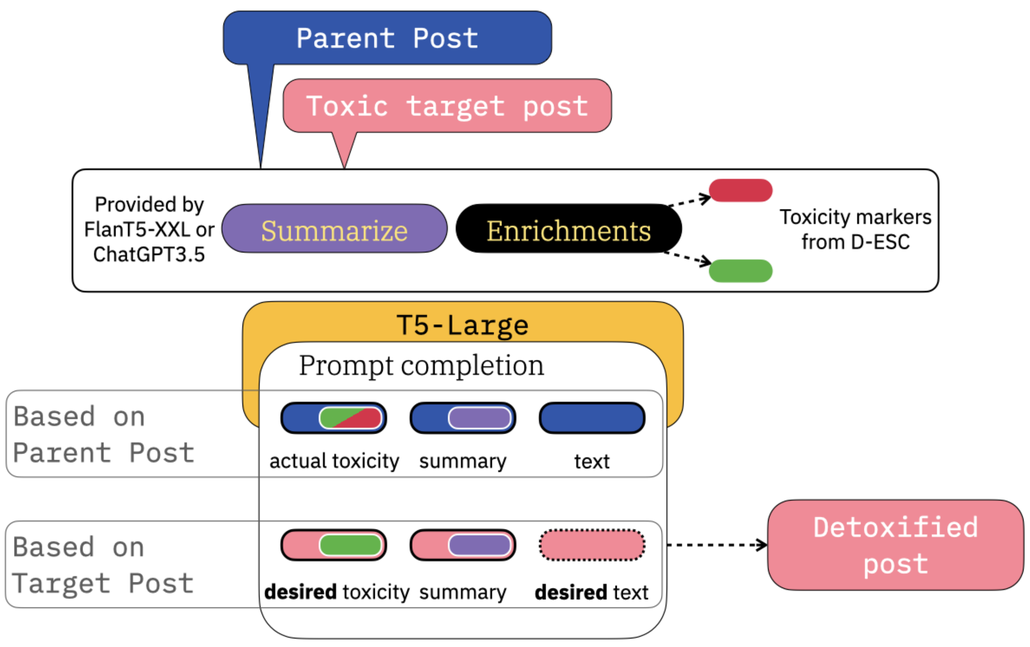

Figure 1: System architecture.

Main Objectives

1. Identify toxicity based on **community standards** (disagree vs. toxic)
2. **Detoxify** text while maintaining author **content** and general **style**

Key Takeaways

1. Generated prompt-completion style tasks using detected toxicity and summaries
2. Summaries generated by either text summarization model or ChatGPT
3. Prompt-completion performed by our fine-tuned T5-Large[1] instance

Motivating example

In this sample Reddit post, the author's underlying message is that "there is no shame in abortion". The toxicity in this post extends from harsh or personal language.

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 post can be rewritten to express the same view without the toxicity:

I have had two abortions and stand by my choices. Your attempts to shame me are ineffective, my friend. It is not productive to shame 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.

Methods

The full system architecture (Figure 1) consists a prompt-generation component which takes any known toxicity marker and summaries of posts and then uses a fine-tuned `T5-Large` instance to generate the final post. Summaries are generated using either a fine-tuned instance of `FLAN-T5-xxl`[2] or ChatGPT[3].

Results

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**. Annotators rated 300 instances divided across the three configurations, with each instance rated by two annotators.

Inter-annotator agreement was calculated using Krippendorff's Alpha with the interval metric, shown in Table 2. PCTS+ChatGPT provides the highest *constructivity* rating – likely due to the built-in moderation framework. However, PCS and PCTS are rated as more *authentic* as they explicitly capture community style (Table 1).

Fluency (fl): comprehensibility and cohesiveness of the generated rewording;

Toxic/constructive (tox/con): degree to which the rewording could contribute to hostility or foster constructive dialogue;

Authenticity/humanness (auth): how natural/human-like the rewording feels;

Relevance (rel): appropriateness of the rewording in relation to the parent post.

	fl	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 1: The results of annotations. 5 annotators rated output posts for fluency, toxicity/constructiveness, authenticity, and relevance.

fl	tox/con	auth	rel
0.46	0.44	0.35	0.44

Table 2: Inter-annotator agreement calculated as Krippendorff's Alpha with interval metric

Next Steps

1. Use deeper semantic representations to guide generation and retain meaning
2. Train to modify other emotions and moral foundations

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 authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. Additional thanks to Reagan Keeney, Adam Brohl, Marilyn Perera, Kyona Jarrett, and Vera Kazakova.

System	Output
Parent	*People are not incubators!*
Original Child	*Indeed. These anti-abortion bills are horribly sexist.*
PCS	*These anti-abortion bills are gonna make us all sick.*
High Tox PCT	*I'm not a person, but I am an incubator.*
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 3: Example System outputs for a parent-child post pairs.

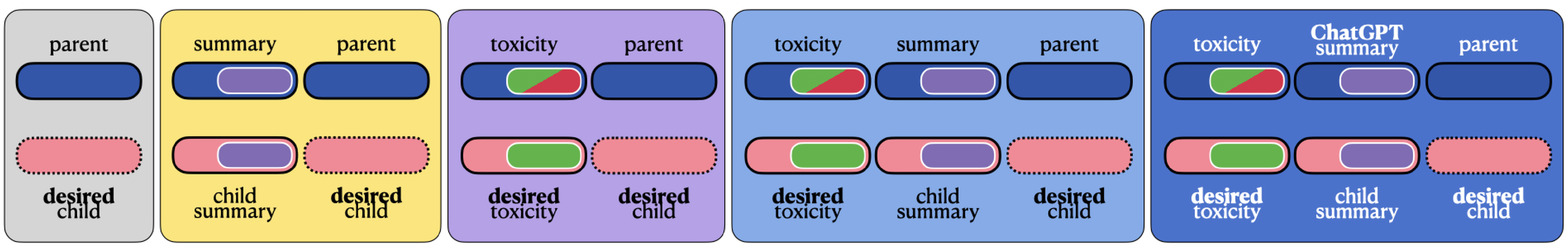

Figure 2: Variants of the prompt generation system. Each one takes in a parent post and a target child post. At training time, the toxicity markers are extracted from the text. At generation time, the child toxicity marker is set to 'low toxicity'.

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

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

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