RoboCop:
Remediation Of Biased Observations and Commentary for Optimal Politness
A Reinforcement Learning Approach to Detoxifying Hate Speech

Luca Carroz, Baptiste Maquignaz, Xavier Ogay

Group 3

Problem definition

Existing detoxification models, such as Paradetox and GreenLlama, struggle to maintain the original semantic meaning of toxic sentences while removing offensive content. This challenge arises because detoxified translations often differ only minimally from the toxic originals, leading to trivial solutions where models merely copy input sentences with minor changes. To address this, reinforcement learning was explored to train a BART model to preserve strong semantic and contextual meaning during detoxification. Despite efforts, models tend to converge on simple copying behaviors, necessitating prolonged training. Incorporating a toxicity classifier aimed to penalize harmful outputs and prevent naive copying, but meaning preservation remains a significant obstacle. This project seeks to refine detoxification approaches to achieve high fidelity in meaning retention while effectively neutralizing toxicity.

Key Related Works

Early online detoxification models treated the task as a style transfer problem, often limited by the lack of parallel datasets.

Paradetox, released in 2022, addressed this by creating a dataset of 10,000 non-toxic paraphrases, enabling effective supervised training with a fine-tuned BART model over 10,000 epochs.

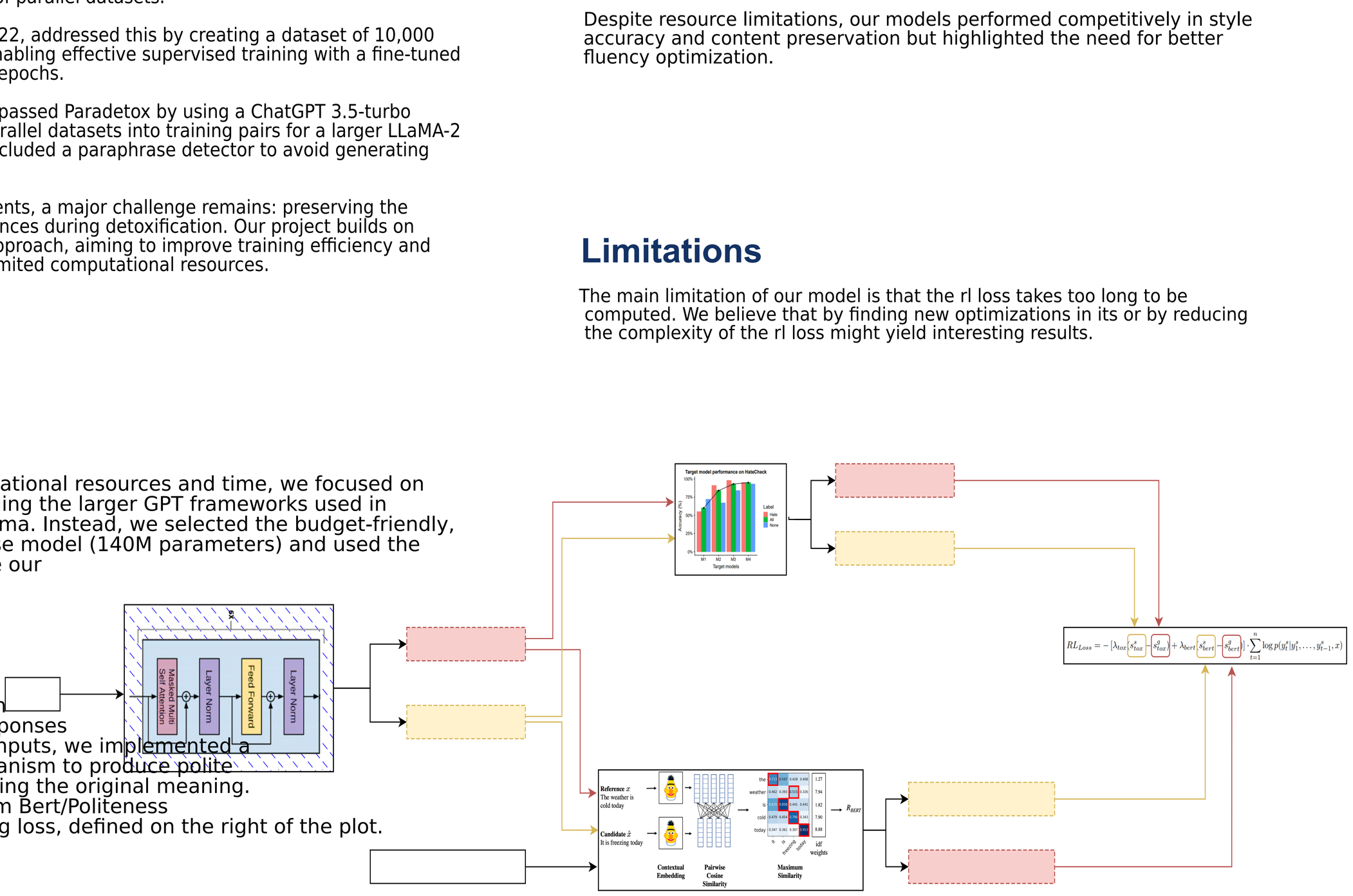
In 2024, a new model surpassed Paradetox by using a ChatGPT 3.5-turbo pipeline to convert nonparallel datasets into training pairs for a larger LLaMA-2 model. This model also included a paraphrase detector to avoid generating harmful content.

Despite these advancements, a major challenge remains: preserving the original meaning of sentences during detoxification. Our project builds on Paradetox's fine-tuning approach, aiming to improve training efficiency and meaning retention with limited computational resources.

Method

Due to limited computational resources and time, we focused on smaller models, avoiding the larger GPT frameworks used in projects like GreenLlama. Instead, we selected the budget-friendly, fine-tunable BART-base model (140M parameters) and used the PEFT library to reduce our training load.

Inspired by GreenLlama's findings, which noted that large l anguage models often default to generic responses when handling toxic inputs, we implemented a reinforcing loss mechanism to produce polite outputs while preserving the original meaning. This involved a custom Bert/Politeness reinforcement learning loss, defined on the right of the plot.



Dataset(s)

We used the dataset from Paradetox [2].
This dataset consists of toxic sentences, together with their non-toxic rewriting(s).
For each toxic input, there can be up to 3 non-toxic rewritings.

MODEL	BLEU	STA	SIM	FL	J
PARADETOX	64.53	0.89	0.86	0.89	0.68
$Ours_1$	47.56	0.71	0.75	0.35	0.19
$Ours_2$	43.32	0.87	0.68	0.17	0.10
$Ours_3$	46.78	0.73	0.76	0.33	0.18

Table 1. EVALUATION OF DETOXIFICATION MODELS

Validation

To validate our model, we first trained it for 100 epochs without reinforcement loss to compare with the Paradetox baseline. This initial model showed good style accuracy and content similarity but had poor fluency and BLEU scores. Short training allowed the model to effectively remove toxicity without significantly altering sentence meaning, but it struggled with grammatical correctness.

Next, we extended training to 1000 epochs, expecting improved fluency. However, the model began overfitting after only 100 epochs, likely due to sub-optimal hyperparameters. Fine-tuning these parameters was challenging due to long training times (approximately 12 hours per run on a T4 GPU).

Our third model trained for 100 epochs without reinforcement loss and an additional 40 epochs with reinforcement loss. The reinforcement loss significantly slowed down training (by a factor of 50), and its impact was minimal, potentially due to the BART model's inability to generate output tokens sequentially, leading to less coherent sentences.

Table 1 summarizes our results compared to Paradetox. We evaluated models on BLEU score, style accuracy (STA), content preservation (SIM), and fluency (FL), with a joint metric (J) calculated as the product of STA, SIM, and FL. While our models achieved comparable style accuracy and content preservation, they lagged in fluency. Unlike Paradetox, we did not manually filter our test set, which may have slightly impacted our metrics.

Despite resource limitations, our models performed competitively in style accuracy and content preservation but highlighted the need for better fluency optimization.

Limitations

The main limitation of our model is that the rl loss takes too long to be computed. We believe that by finding new optimizations in its or by reducing the complexity of the rl loss might yield interesting results.

Conclusion

This study explored reinforcement learning to preserve semantic integrity while detoxifying online content. By integrating a BART model with a toxicity classifier and BERTScore-based reinforcement, we aimed to balance content fidelity and non-offensiveness. Despite limited computational resources, we hope that further work will be done to the rl loss to finally achieve notable detoxification with enhanced meaning preservation.

References

[1] C. N. d. Santos, I. Melnyk, and I. Padhi, “Fighting offensive language on social media with unsupervised text style transfer,” arXiv preprint arXiv:1805.07685.
[2] 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), pp. 6804–6818, 2022.
[3] M. T. I. Khondaker, M. Abdul-Mageed, and L. V. Lakshmanan, “Greenllama: A framework for detoxification with explanations,” arXiv preprint arXiv:2402.15951, 2024