Ensuring Safety in Summarization: Evaluating Toxicity Suppression in Abstractive Summarization

Ahsan Kabir Nuhel, Arunima Chaurasia, Keerthan Shekarappa

May 5, 2025

1 Introduction

Large Language Models (LLMs) have revolutionized the field of automatic summarization, offering fluent and coherent summaries across various domains. However, when tasked with summarizing documents that contain toxic, hateful, or biased language, the challenge becomes twofold: retaining essential information while suppressing harmful content.

This research explores how well LLMs suppress toxicity during summarization. Rather than focusing on whether they introduce new toxicity, our emphasis is on whether they can act as detoxifiers, reducing or removing toxicity present in the source document while preserving meaning. This is critical for deploying LLMs in environments such as media monitoring, social platforms, and customer feedback analysis.

Recent research has addressed toxicity in LLMs, but mostly outside the context of summarization. Gehman et al. [1] introduced the RealToxicityPrompts dataset to evaluate how language models like GPT-2 produce toxic content. In the summarization domain, Krishna et al. [2] and Fabbri et al. [3] emphasize hallucination and quality evaluation, respectively, without incorporating toxicity assessment. Kaur et al. [4] take a step toward fairness in summarization, applying bias mitigation on the CNN/DailyMail dataset with limited attention to toxicity. Although Narayan et al. [5] introduced the Multi-News dataset for multi-document summarization, none of the cited studies have used it to evaluate toxicity handling in LLM-generated summaries.

1.1 Research Questions

- **RQ1:** Can LLMs effectively reduce or suppress toxic content present in source documents during summarization?
- **RQ2:** Which LLMs, such as BART [6] and T5 [7], demonstrate the highest effectiveness in mitigating toxicity?
- **RQ3:** How effectively do the generated summaries preserve the original content?

2 Methodology

2.1 Technical Approach

1. **Dataset:** We used the Multi-News dataset [8], which contains sets of news articles and corresponding human-written summaries for multi-document summarization tasks.
2. **Preprocessing:** The dataset underwent basic preprocessing including lowercasing, removal of special characters, and whitespace normalization. Token cleaning and input formatting were performed using standard NLP tokenizers.
3. **Toxicity Detection (Input):** The original documents were scored using the Perspective API [9] and Detoxify [10] to establish baseline toxicity levels.
4. **Summarization Models:** Summaries were generated using two pretrained models: BART (facebook/bart-large-cnn) and T5 (t5-base), without further fine-tuning.

5. **Toxicity Detection (Summaries):** Generated summaries were evaluated for toxicity using the same classifiers to assess suppression effectiveness.
6. **Content Evaluation:** ROUGE-1, ROUGE-2, and ROUGE-L [11] scores were computed against reference summaries to evaluate content preservation.
7. **Analysis:** The toxicity levels and content scores before and after summarization were compared to understand the effectiveness of the models in mitigating toxicity while retaining relevance.

2.2 NLP Techniques

- **Pre-trained Models:** Utilization of pre-trained models such as BART [6] and T5 [7] for abstractive summarization tasks.
- **Preprocessing Techniques:** Application of preprocessing techniques including tokenization, stopword removal, and stemming to normalize the input data.
- **Toxicity Detection:** Utilization of toxicity detection models such as Perspective API and Detoxify to evaluate the toxicity levels of both source documents and generated summaries.
- **Content Evaluation Metrics:** Use of metrics such as ROUGE-1, ROUGE-2, and ROUGE-L to evaluate the content preservation and quality of generated summaries.

3 Team Contributions

3.1 Shared Responsibilities

All Members: Data collection, prompt design, code review, and final report/poster preparation.
Deliverables:

- Public GitHub repository for the project is available here.
- Poster summarizing methodology and findings.

3.2 Individual Responsibilities

- **Arunima Chaurasia**
  - **Role:** Dataset creation, including curation of toxic content and formatting for summarization.
  - **Deliverables:** Cleaned and labeled dataset with source-summary toxicity annotations.
- **Keerthan Shekarappa**
  - **Role:** Model prompting and summarization experiments across multiple LLMs.
  - **Deliverables:** Set of model-generated summaries and prompt templates.
- **Ahsan Kabir Nuhel**
  - **Role:** Toxicity scoring, evaluation metrics, and comparative analysis framework.
  - **Deliverables:** Evaluation script and analysis report.

4 Evaluation and Dataset

4.1 Dataset Description

We will be using Multi-News dataset [8] for our use case.

The Multi-News dataset comprises news articles and their corresponding human-generated summaries sourced from newser.com. These summaries are crafted by professional editors and provide links to the original articles referenced.

Field	Description	Type
document	Text of news articles separated by special token "|||||"	string
summary	News summary	string

4.2 Experimental Setup

We will evaluate using the following metrics on our train / validation / test split for each model.

Toxicity Suppression Metrics:

- Pre- and post-summary toxicity score delta (using Perspective API) [9] or Detoxify [10]
- Retained content score (using BERTScore or ROUGE-1, ROUGE-2, and ROUGE-L) [11]

References

[1] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "Realtoxicityprompts: Evaluating neural toxic degeneration in language models," *Findings of EMNLP*, 2020.

[2] K. Krishna, Z. Zhang, and M. Iyyer, "Halueval: Benchmarking the faithfulness of language models in multi-document summarization," in *EMNLP*, 2021.

[3] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, "Summeval: Re-evaluating summarization evaluation," in *NAACL*, 2021.

[4] T. Kaur, S. Krishna, R. Shah, A. Sangwan, L. Subramaniam, and A. Sharma, "Mitigating social bias in text summarization models," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

[5] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization," *EMNLP*, 2018.

[6] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[7] 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.

[8] A. R. Fabbri, I. Li, T. She, S. Li, and D. R. Radev, "Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model," 2019.

[9] Jigsaw and Google, "Perspective api." https://www.perspectiveapi.com/, 2017. Accessed: 2025-05-03.

[10] L. Hanu and Unitary team, "Detoxify." Github. https://github.com/unitaryai/detoxify, 2020.

[11] M. Barbella and G. Tortora, "Rouge metric evaluation for text summarization techniques," *Available at SSRN 4120317*, 2022.