Detoxigram: Helping users understand why content is classified as toxic

Joaquín Navajas Emmanuel Iarussi Santiago Corley
Alexia Aquino Andrés Cotton Luz Alba Posse

Torcuato Di Tella University

May 30, 2024

1 What Have We Learned?

1.1 What We Learned About Toxicity

1.1.1 How to Define It

Inspired by previous literature [1, 3, 5], we characterized toxicity as a range of disrespectful and harmful user-generated content, not as a binary element. As such, we decided to measure and evaluate toxicity using a scale that aligns with how individuals might perceive and react towards toxic messages online. The scale we used is the following:

- **Non-toxic:** Respectful and constructive. Does not contain personal attacks or offensive content.
- **Slightly Toxic:** Although mostly respectful, it suggests a lack of appreciation for the viewpoints of others. It does not directly attack individuals or groups.
- **Moderately Toxic:** Features a clearly disrespectful tone. Does not involve violent attacks.
- **Highly Toxic:** Insulting and aggressive. Targets groups and individuals based on their gender, ethnicity, sexual orientation, ideology, or religion.
- **Extremely Toxic:** In addition to the elements of highly toxic messages, this includes threats or calls to violent action.

In accordance with this definition, an online survey (N=300) was conducted with a sample of American participants, balanced by gender and political orientation (167 female, aged 40.69 $\pm$ 11.95 yr). The survey's primary objective was to assess individual perceptions of message toxicity within various Telegram

channels. They were recruited through Prolific, an online research platform, and invited to complete the survey via SurveyMonkey. They were informed that their participation was completely voluntary and that they could withdraw their participation at any time without penalty.

Participants rated the toxicity of 30 messages (sourced from different Telegram channels) on the 5-point scale (0-4) mentioned before. Attention checks were incorporated to ensure data quality, and participants failing two or more of them were excluded from compensation. A £0.90 incentive was provided upon successful completion to encourage engagement and maintain data quality. We used the recollected data as ground truth for our classifier.

1.2 Multiple Attributes in Toxicity

An observation from the literature is that polarizing or toxic content has different attributes that are key to understanding unhealthy online conversations [4]. We believe that both detecting and communicating these attributes are crucial components of a toxicity classifier. To help build this classifier, we used the UCC dataset, a rich and annotated set of online comments.

Analyzing this dataset, we realized some of the labeled attributes were similar and, as such, redundant. Using Principal Components Analysis, we determined that the most important attributes to detect are: sarcasm, antagonism, generalization, and dismissiveness.

1.3 How to Communicate It

As far as we know, there are no other toxicity classifiers that combine accurate and nuanced metrics with tailored and persuasive explanations of their classifications. Moreover, they do not showcase to the user healthier alternatives of communication.

1.4 What We Learned About the Technology

In our literature overview of existing toxicity detection models, we concluded that different technologies have proven accurate for the task. The two main ones are outlined below:

- When prompted effectively, generative LLMs are accurate classifiers for nuanced language tasks and can generate persuasive messages. However, they can be quite expensive, slow, and memory intensive.
- On the other hand, BERT models are often cheap to train and deploy and accurate, but they may struggle with nuanced classifications.

We concluded that leveraging these two classification tools can yield great results. We propose a classification pipeline that integrates these classifiers with the Telegram API to develop agents that interact smoothly with users. This setup makes it easy to fetch messages from public Telegram channels while preserving user privacy, as we do not know who sent the messages.

2 Choosing the Right Generative LLM

The first question we had to answer when choosing the generative LLM for this project was: Are we aiming to deploy a full model for this project, or do we prefer using external services provided by companies, such as OpenAI, Anthropic, or MistralAI?

We first explored a very efficient implementation of an LLM, namely llama.cpp. This tool allows us to run inference on CPU, reducing costs of deployment. However, running our models locally was still computationally expensive. Therefore, we decided to go in a different direction to build the minimum viable product. To keep the energy demand of our pipeline low, we restricted our search to relatively small LLMs that can generate high-quality responses through an API. More importantly, we wanted our model to be open source.

We decided to use Mixtral of Experts [2], an LLM from Mistral AI, which we found to be the best option with respect to cost and performance trade-offs. Choosing this model also allows us to potentially fine-tune it or deploy it locally if we ever decide to.

3 What Have We Built?

We have built Detoxigram, a tool that analyzes toxicity in text channels, explains its decisions, and detoxifies messages with the goal of reducing negative interactions and promoting constructive dialogue. As far as we know, Detoxigram is the first-ever toxicity classifier that also accurately explains its classifications. Detoxigram both classifies Telegram channels and provides metrics with tailored explanations of its classification.

To achieve this, we proposed a toxicity-detection pipeline that combines generative LLMs and BERT to exploit their strengths efficiently.

We decided to use the BERT-based classifiers for two different tasks:

- Coarse-grained toxicity detection of Telegram messages.
- Detecting unhealthy-speech attributes of the Telegram messages.

The purpose of the first task is to minimize the volume of messages being inputted into the LLM. By choosing to input to the LLM only those messages deemed the most toxic by our BERT classifier, we are able to explain the toxicity of a Telegram channel through a low resource approach without sacrificing nuance. For this task, we used a pre-trained BERT classifier, trained on the Toxigen dataset.

The purpose of the second task is to accurately demonstrate to users the various dimensions of unhealthy speech that a channel might exhibit. By providing specific information about the different dimensions of toxicity in Telegram messages, users can make more confident and informed decisions about the content they consume. For this task, we trained a BERT-based classifier using the dataset mentioned in the previous section.

We decided to use generative LLMs for the following tasks:

- Fine-grained toxicity detection of Telegram messages.
- Explaining to the user the reasons behind each classification.
- Detoxifying toxic messages.

The Detoxigram classification pipeline can be outlined through the following steps:

1. We extract the last 50 messages of a given Telegram channel.
2. We classify these messages using a BERT-based toxicity classifier.
3. We then take the 10 most toxic messages from the previous step and classify them again using a generative LLM. This ensures we accurately portray the toxicity of the Telegram channel.
4. Finally, we prompt the generative LLM to produce a short text explaining the classification of the Telegram channel, and we use the second BERT-based classifier to detect the toxicity attributes of the channel.

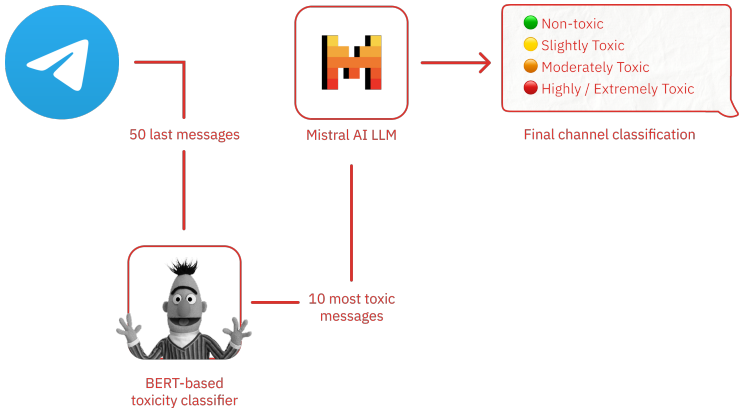

Figure 1: Classification Pipeline

4 What Users Can Do with Detoxigram?

Telegram users who use Detoxigram have the following tools at their disposal to manage and understand toxicity in their conversations:

- **Analyze a Channel:** Users can use Detoxigram to analyze the toxicity level of any public Telegram channel. By simply inputting the channel name, our tool will scan and evaluate recent messages to provide a classification. This helps users understand the general tone and health of the conversations within the channel.

- **Understand Why It Is Toxic:** Detoxigram not only identifies toxic messages but also provides detailed explanations for each classification. This feature helps users understand the reasons behind the toxicity ratings, offering insights into the specific elements and language that contribute to harmful interactions, while also explaining the consequences for the user.
- **Visualize the Dimensions of Toxicity:** Our tool categorizes toxicity into various dimensions, such as sarcasm, antagonism, generalization, and dismissiveness. Users will be able to visualize these dimensions to get a nuanced view of how toxicity manifests in different forms.
- **Detoxify Your Own Messages:** Detoxigram also offers a unique feature that allows users to detoxify their own messages before posting them. By running their text through our tool, they can receive suggestions on how to rephrase or adjust their message to avoid potential toxicity. This proactive approach helps users to communicate more effectively and constructively, fostering healthier interactions.

5 Main Advantages of Detoxigram

5.1 Tailored and Persuasive Explanation

Detoxigram is able to provide an explanation of why the channel was classified with a certain level of toxicity, a summary of the main topics discussed in the channel, and a brief overview of the potential consequences for the user engaging with such content.

5.2 Costs

By using a BERT classifier to reduce the number of messages classified by our generative LLM, we reduce in 5 the amount of calls done to the API of the chosen LLM. We calculated these values and concluded that the cost of analyzing a channel using our pipeline is, on average, $0.0017 USD.

5.3 Low Energy Consumption

By leveraging a BERT model and a generative LLM, we ensure that our pipeline is not as resource-intensive as it would be when only using a generative LLM. The BERT model we use has only 110 million parameters. A relatively small generative LLM (for example, LLaMA 7B) has 60 times more parameters.

5.4 Behavioral Design Based UX

As a Telegram bot, Detoxigram is seamlessly integrated into the interface the users employ on a daily basis. It requires no downloads or installations. Information is provided to the user gradually according to its requests. Our classifications are communicated in combination with an intuitive color code that

works as a "traffic light" that synthesizes the classification. Toxicity dimensions are shown with an image that is coordinated with the color code matching the overall classification of the channel.

5.5 Open Source

Developed with open source technology, Detoxigram doesn't have third-party dependency and provides greater modular flexibility for further developments and updates.

6 What Are Our Plans for the Future?

- **Twitter Analyzer (Live beta):** We've developed a real-time Twitter data analysis tool that provides detailed insights into the level of toxicity in a user's content. This helps to understand the implications of the content we consume in a very engaging way.
- **Twitter Bot Integration:** We aim to create a bot seamlessly integrated into Twitter, providing users with an intuitive way to interact with our platform, perform various tasks, and receive personalized recommendations on the content they consume.
- **Fine-tuning of an LLM:** To enhance the performance of our toxicity detection models, we're exploring the fine-tuning of a smaller LLM specifically for classification tasks in the context of text toxicity, aiming to achieve greater accuracy and efficiency.
- **Scalability Enhancements:** Recognizing the growing demand for our services, we're committed to scaling our infrastructure to support thousands of concurrent users. This involves optimizing our code, deploying load balancers, and leveraging cloud-based services to ensure high availability and performance.
- **Expanding to WhatsApp:** In addition to Twitter, we're considering deploying our functionalities into WhatsApp, enabling users to access our services more conveniently through a widely used messaging platform.
- **Explore Improving the Performance of the LLM Classifier:** To enhance the performance of our toxicity detection models, we're exploring the fine-tuning of a smaller LLM specifically for classification tasks in the context of text toxicity, aiming to achieve greater accuracy and efficiency.
- **API Development:** Detoxigram API will empower developers to integrate our toxicity detection and moderation capabilities into their own applications.

References

[1] Paula Fortuna, Juan Soler-Company, and Leo Wanner. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524, 2021.

[2] Albert Q. Jiang et al. Mixtral of experts. *arXiv preprint arXiv:2202.12345*, 2024.

[3] Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4262–4274. Association for Computational Linguistics, 2021.

[4] Ilan Price, Jordan Gifford-Moore, Jory Flemming, Saul Musker, Maayan Roichman, Guillaume Sylvain, Nithum Thain, Lucas Dixon, and Jeffrey Sorensen. Six attributes of unhealthy conversations. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 114–124. Association for Computational Linguistics, 2020.

[5] Zhixue Zhao, Ziqi Zhang, and Frank Hopfgartner. A comparative study of using pre-trained language models for toxic comment classification. *Companion Proceedings of the Web Conference 2021*, 2021.