Detecting and Mitigating Toxic Comments in Online Platforms Using Toxic Bert and Topic Modeling

Group #1

Marina Mitiaeva, Mervin McDougall, Lakshmi Peram, Nick DeVita

Overview

Reddit’s current moderation system employs a multi-layered approach that integrates automated tools, human review, and user reports. While the platform claims to be “constantly building and refining tools to proactively identify content and behavior that violates policies” [1], the application of natural language processing (NLP) remains limited. Recent advancements, such as the use of large language models (LLMs) for harassment detection, have only begun to emerge, coinciding with the company’s IPO efforts [2]. Despite these developments, a significant portion of Reddit’s moderation—approximately two-thirds—relies on manual intervention, placing a heavy burden on unpaid volunteer moderators. This dependency on volunteer labor exposes the platform to vulnerabilities, as highlighted by the 2023 protests, where moderators opposed Reddit’s implementation of new data access fees for third-party app developers [3].

The goal of this project is to enhance Reddit’s moderation processes by proposing a proactive framework capable of identifying high-risk topics that are more likely to attract toxic, viral comments. Our analysis focuses on a large Reddit dataset [4] containing questions, scores, and their top-voted answers to identify highly engaging topics and evaluate their relationship with title toxicity and user engagement. To achieve these objectives, we employed a combination of descriptive statistics, word cloud visualizations for exploratory data analysis, unsupervised learning techniques to uncover hidden topics through Latent Dirichlet Allocation (LDA) modeling [5], and a pre-trained Toxic-BERT model [6] for toxicity prediction.

The findings were both insightful and unexpected. While the most popular topics—primarily centered on politics, technology, and money/taxes—garnered significant engagement, their corresponding toxicity scores were surprisingly low. This suggests that while these topics are often highly engaging, they do not necessarily drive harmful discussions within the sampled dataset. These results lay the groundwork for developing a more advanced, fine-tuned model capable of analyzing thread patterns and estimating the potential toxicity of topics in real time. Such a system would enable moderators to proactively address harmful content, reduce manual workloads, and promote healthier, more constructive interactions within Reddit’s communities.

Prediction Inference

The primary objective of this research is to quantify the toxicity of Reddit posts and examine its relationship with thread topics and key engagement metrics. This study aims to address the following research questions:

- **RQ1:** What are the most viral topics on Reddit?
- **RQ2:** What is the relationship between toxicity of the title, title length, and user engagement on Reddit?

To address these questions, we utilized Latent Dirichlet Allocation (LDA) [5] to identify the key topics of the threads and employed a pre-trained Toxic-BERT [6] model to quantify the toxicity scores of the titles. Subsequently, we analyzed the relationships between the identified topics, title toxicity, and engagement metrics. This analysis revealed several key insights:

- **Viral topics:** The analysis of viral Reddit topics reveals a strong dominance of politics, technology, and personal narratives, with keywords such as "government," "vote," "data," and "steam" driving engagement. There is limited diversity in topics, as nearly half of the posts are controlled by just two dominant themes out of the 20 detected through LDA modeling. This suggests a concentrated focus on a few key interests within the platform's user base. Notably, there is no clear pattern between thread length and virality, indicating that content relevance and relatability play a greater role in driving engagement than the depth of discussion.

- **Title Toxicity, Title Length, and Engagement:** The analysis of title toxicity revealed very low toxicity levels across the dataset. There is only one topic in the top 10 viral threads has been detected with high probability, while being quite clear using explicit language. Correlation analysis indicated negligible relationships between toxicity and other features. Specifically, there was a slight negative correlation between toxicity scores and engagement scores (-0.073), suggesting that higher toxicity titles might slightly reduce user engagement. Additionally, the correlation between toxicity scores and title length was almost non-existent (-0.032), while the weak positive correlation between title length and engagement scores (0.069) suggests minimal dependency, indicating these features operate largely independently.

These findings suggest that while topics play a significant role in driving engagement, toxicity levels are surprisingly low in the sampled dataset and do not appear to substantially influence user behavior. The lack of strong relationships between toxicity, title length, and engagement underscores the complexity of Reddit's content dynamics and highlights the need for more nuanced models to capture the interplay of these features.

Data

In our project, we utilized a sample of the reddit_question_best_answers dataset, released in July 2022 and hosted on the Hugging Face data repository [4]. The dataset comprises 1,833,556 posts collected between January 2010 and June 2021. Each entry includes the title of the post, the main body of text, the score for the post, and a nested structure of associated answers with their respective scores. Due to technical limitations and memory restrictions in Google Colab, we sampled 50,000 observations for analysis to ensure computational feasibility.

Table 1 offers a detailed overview of the descriptive statistics for the sampled dataset, uncovering patterns and relationships between key features. By combining textual and engagement metrics, we aimed to better understand the dynamics of user activity and the factors contributing to meaningful discussions.

	Mean	**Std**	**25%**	**50%**	**75%**
Question score	153.4522	403.9283	10	26	91
Answers Count per question	4.7376	8.9857	1	2	5
Answers score per post	13.7665	49.7303	3.6	5	7.75
Title length	70.9052	39.7671	43	61	88
Body length	930.0921	1171.7703	279	532	1043
Answers length	484.2681	608.2985	166	282	552
Combined answers length	2295.7543	6177.0468	385	946	2254
Thread length	3154.3103	6363.4522	943	1732	3339

Table 1. Descriptive statistics of Reddit dataset

In our analysis, we primarily focused on the textual features (title, body, answers) and user engagement metrics, such as question and answer scores. To enhance our understanding of Reddit's Q&A ecosystem, we introduced a new feature—thread—which consolidates all textual components (title, body, and answers) for each post into a single, comprehensive representation.

The question scores show considerable variability, with a mean of 153.45 and a standard deviation of 403.93, reflecting a highly skewed distribution. While most questions receive modest engagement, a small subset achieves exceptional popularity, likely driven by their relevance, clarity, or alignment with trending topics. This is further highlighted by the large gap between the 50th percentile (26) and the 75th percentile (91). The relatively low engagement at the 25th percentile (10) suggests that many questions fail to resonate with the community, potentially due to vague phrasing, insufficient detail, or niche appeal. This pattern aligns with the power-law nature of online interactions, where a few posts dominate attention.

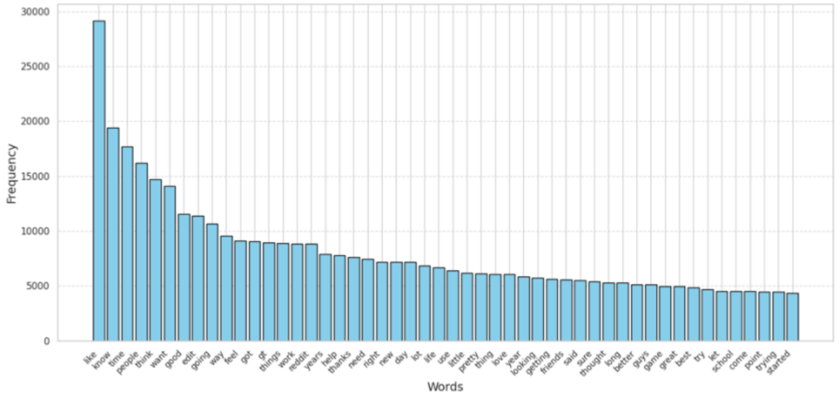
The answer count per question further illustrates this asymmetry. Although the mean is 4.73, the median of 2 indicates that most questions receive limited responses, with only a minority sparking extensive discussions. The high standard deviation (8.99) reinforces the variability in community engagement, where a small fraction of questions attracts significantly more interaction.

Comparing textual features, thread length, averaging 3154.31 characters, provides a comprehensive measure of engagement by encapsulating the combined textual effort of the original poster and the community's responses. This aggregated feature offers a holistic perspective on user interaction, capturing both the initial question's quality and the depth of collective discourse it generates. Unlike isolated metrics such as question score or body length, thread length reveals the collaborative nature of Reddit's ecosystem, making it a robust indicator for deeper analysis. By focusing on threads, we can better explore the dynamics of topics and engagement, identify factors that drive meaningful discussions, and characterize the attributes of toxic conversations.

Key insights and unexpected resutls

During our initial experiments with topic modeling, we encountered challenges caused by the presence of numerous stop words, which negatively impacted both the modeling process and the interpretability of the results. To address this issue, we tested various stop word corpora and ultimately found the most effective solution to be a combination of stop words from SpaCy, NLTK, and Scikit-learn, supplemented by a custom list tailored to our dataset. This approach significantly improved the quality of the tokenized text by reducing noise. Figure 1 illustrates the 50 most common words in the dataset after applying tokenization, lowercasing, stop word removal, and the cleaning of non-alphabetical characters.

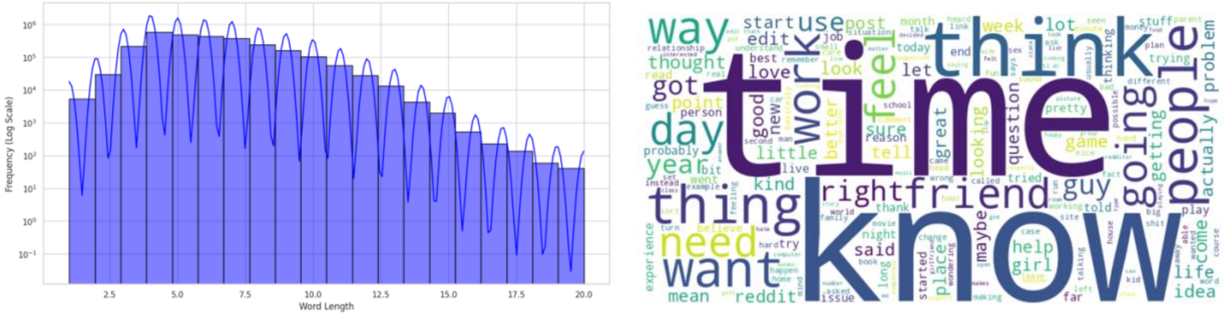
Figure 1. Frequency Distribution of the Top 50 Most Common Words in the Corpus



While many of the expected functional words (such as common verbs) appeared, there were some surprising results. For instance, the word 'time' emerged as a frequent term, which could potentially hold meaningful insights if we had metadata about the publication times of posts, but this is not the case. The 50 most common words collectively appeared 405,539 times in the corpus of 2,719,348 tokens, accounting for approximately 15% of the total text.

Another intriguing finding was the appearance of the term 'gt' among the top 15 most frequent words. This prompted us to examine the distribution of word lengths within the corpus. The analysis revealed that short words make up a substantial proportion of the vocabulary and removing them could result in a significant loss of information. This observation was further confirmed by a word cloud, which highlighted meaningful short words like 'sex' and 'ai'. These results underscored the need to carefully balance noise reduction with the preservation of informative terms, even when they are short in length.

Figure 2. (a) Distribution of Word Lengths and (b) Word Cloud Representation of the Corpus



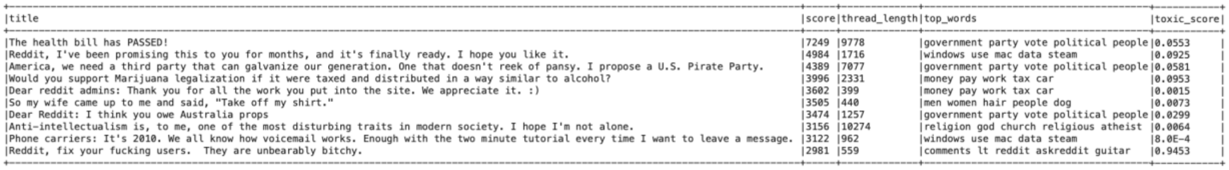
a

b

Overall, our corpus consisted of 2,708,135 tokens, with only 78,984 unique words. This means that just 2.92% of the words in the corpus are unique, highlighting a high degree of repetition and concentration within the text. These findings reflect the challenges and opportunities inherent in processing and analyzing a dataset of this scale.

The LDA results revealed a lack of diversity in the topics of the top 10 threads, which predominantly focus on political issues, technology, and money-related discussions (Figure 3). Recurring keywords such as "government," "party," and "tax" indicate that threads addressing societal and financial concerns drive the most engagement, likely due to their polarizing and discussion-provoking nature. Despite this high engagement, the toxicity levels in these threads were surprisingly low, suggesting that even in discussions of potentially controversial topics, the titles remain relatively neutral and do not substantially contribute to toxic discourse.

Figure 3. Topic Modeling Results for the Top 10 Most Engaged Threads with Title Toxicity Scores



```
+-------------------------------------------------------------------------------------------------------------------------------------+-----+-------------+--------------------------------------+-----------+
|title                                                                                                                                |score|thread_length|top_words                             |toxic_score|
+-------------------------------------------------------------------------------------------------------------------------------------+-----+-------------+--------------------------------------+-----------+
|The health bill has PASSED!                                                                                                          |7249 |9778         |government party vote political people|0.0553     |
|Reddit, I've been promising this to you for months, and it's finally ready. I hope you like it.                                      |4984 |1716         |windows use mac data steam            |0.0925     |
|America, we need a third party that can galvanize our generation. One that doesn't reek of pansy. I propose a U.S. Pirate Party.     |4389 |7077         |government party vote political people|0.0581     |
|Would you support Marijuana legalization if it were taxed and distributed in a way similar to alcohol?                               |3996 |2331         |money pay work tax car                |0.0953     |
|Dear reddit admins: Thank you for all the work you put into the site. We appreciate it. :)                                           |3602 |399          |money pay work tax car                |0.0015     |
|So my wife came up to me and said, "Take off my shirt."                                                                              |3505 |440          |men women hair people dog             |0.0073     |
|Dear Reddit: I think you owe Australia props                                                                                         |3474 |1257         |government party vote political people|0.0299     |
|Anti-intellectualism is, to me, one of the most disturbing traits in modern society. I hope I'm not alone.                           |3156 |10274        |religion god church religious atheist |0.0064     |
|Phone carriers: It's 2010. We all know how voicemail works. Enough with the two minute tutorial every time I want to leave a message. |3122 |962          |windows use mac data steam            |8.0E-4     |
|Reddit, fix your fucking users.  They are unbearably bitchy.                                                                         |2981 |559          |comments lt reddit askreddit guitar   |0.9453     |
+-------------------------------------------------------------------------------------------------------------------------------------+-----+-------------+--------------------------------------+-----------+
```

Table 2 presents the descriptive statistics for toxicity scores obtained for subsample of the titles, reflecting the probability of toxicity detected in the text, where values closer to one indicate higher toxicity. Surprisingly, the distribution of scores is notably low, with a mean of 0.0206 and a standard deviation of

0.0234. This may suggest that Reddit, at least in this dataset, is not particularly suited for this model, possibly due to the neutrality of the data or a lack of alignment between the model's training data and the specific topics in this dataset, despite the inclusion of civil comments in the model's training corpus [6].

	Mean	**Std**	**25%**	**50%**	**75%**
Toxicity score	0.0206	0.0234	0.004	0.01	0.0303

Table 2. Descriptive Statistics of Title Toxicity Scores

Table 3 presents the correlation analysis between title length, toxicity, and engagement score, revealing weak or negligible relationships among these features. A slight negative correlation between toxicity score and engagement score (-0.073) suggests that higher toxicity is associated with marginally lower user engagement. Additionally, the near-zero correlation between toxicity score and title length (-0.032) and the weak positive correlation between engagement score and title length (0.069) indicate minimal linear dependency, highlighting that these features are largely independent and not strongly predictive of one another.

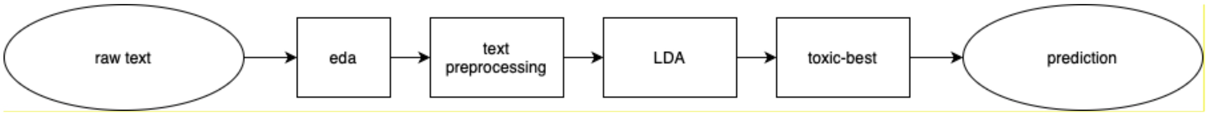
	Toxicity score	Score	Title length
Toxicity score	1	-0.0732165	-0.03165093
Score	-0.0732165	1	0.06945673
Title length	-0.03165093	0.06945673	1

Table 3. Correlation Matrix Between Toxicity Scores, Question Scores, and Title Length

Methodology

In our project, we employed unsupervised learning techniques to uncover hidden topics through topic modeling and utilized a pre-trained model for toxicity prediction. The overall methodology is represented in the project pipeline shown in Figure 3.

Figure 3. Workflow Overview: From Raw Text to Prediction Using Topic Modeling and Toxicity Scoring



After exploring the raw dataset and gaining initial insights, we defined a series of preprocessing steps to prepare the data for analysis. These steps included tokenization, lowercasing, and the removal of stop words using a combination of SpaCy, NLTK, Scikit-learn, and a custom stop word list. Additionally, we cleaned the data by removing non-alphabetical characters and filtering out words with fewer than two characters.

To transform the processed text into a numerical format for topic modeling, we applied a count vectorizer with parameters VocabSize = 20,000 and minDF = 5, ensuring that only terms appearing in at least five documents were included. We then computed the inverse document frequency (IDF) to weight the terms appropriately and defined a Latent Dirichlet Allocation (LDA) model with 20 topics (topics = 20) and a maximum of 10 iterations (maxIter = 10). The output was limited to the five most relevant words per topic, which were mapped back to their corresponding indices. As a result, each document—representing a thread—was tagged or categorized with the keywords generated by the topic modeling process.

Following the topic modeling phase, we employed the pre-trained Toxic-BERT model to calculate a toxicity score for a subsample of the titles and evaluate its relationship with the identified topics. These scores provided a probabilistic estimate of the toxicity level present in the text, with values closer to 1 indicating higher levels of toxicity. This allowed us to assess the alignment between toxic content and specific topics, offering insights into how toxicity varies across different thematic discussions.

To further analyze the interplay between features, we conducted a correlation analysis between question scores, toxicity scores, and text lengths. This step aimed to uncover potential relationships and dependencies among these variables, providing a deeper understanding of the factors influencing user engagement and content dynamics.

Limitations

The challenges encountered during the project can be grouped into four main categories: memory restrictions, library incompatibilities, selective stop-word filtering, and the interpretation of topic modeling results.

One of the primary challenges arose during the initial dataset processing. The dataset, containing 1.8 million observations, exceeded the memory limitations of the free tier for Google Colab. To address this, we sampled 50,000 records for analysis, ensuring computational feasibility while preserving a representative subset for development. Memory constraints also became evident during the Exploratory Data Analysis (EDA) phase, primarily due to the immutable nature of PySpark variables. Intermediate data structures, such as DataFrames, persisted in memory even after processing was complete, leading to frequent memory shortages. To manage this, we explicitly set unused variables to None to free up memory. Additionally, we performed some data transformations within functions, ensuring that variables were automatically destroyed once they were out of scope, thereby optimizing memory usage.

Generating toxicity scores using the pre-trained Toxic-BERT model with PySpark DataFrames presented another significant challenge. Despite creating User Defined Functions (UDFs) to process threads and generate toxicity scores, the process frequently timed out without producing results. This issue is believed to stem from an incompatibility between PySpark and Toxic-BERT. As a workaround, we converted the PySpark DataFrame to a Pandas DataFrame, allowing us to proceed with toxicity scoring. While this approach resolved the issue, it introduced additional computational overhead. Visualizing the findings posed similar challenges, as PySpark data had to be duplicated and converted to formats compatible with visualization libraries such as Matplotlib and WordCloud. This duplication further strained the available memory, limiting our ability to process larger datasets. To mitigate this, we opted for lightweight and simple data structures, such as lists, to pass data between PySpark and other libraries.

For stop-word removal, we combined lists from three well-known libraries—SpaCy, NLTK, and Scikit-learn—along with a custom list to filter out common stop words. Additionally, a rule of thumb was applied to filter out words with fewer than three characters. However, during topic modeling, we discovered that certain short words, such as “God” in a thread titled “The health bill has PASSED!”, carried significant contextual meaning. Removing such words risked losing valuable insights, prompting us to refine our filtering criteria.

Interpreting the results of topic modeling proved to be particularly challenging. While many generated keywords were meaningful, some, such as “lt”, were ambiguous and lacked obvious context, making them difficult to interpret. This highlighted the need for greater contextual understanding to effectively interpret less meaningful terms. Despite these challenges, we refined our methodology to focus on the most interpretable and relevant keywords, ensuring the topic modeling process provided actionable insights.

Discussion

Overall, we successfully achieved our primary goal of obtaining toxicity scores for Reddit threads and analyzing potential drivers of toxicity. However, the unexpectedly low toxicity scores made it challenging to interpret the results and draw actionable conclusions. This limitation suggests that either the dataset's neutrality or the model's alignment with the specific topics may have influenced the outcomes. The project could have been more insightful with the use of an alternative version of the toxicity model that supports more nuanced features, though such models are not currently compatible with PySpark or SparkNLP.

Further refinements in preprocessing could significantly improve the quality of insights derived from the dataset. Incorporating more advanced techniques, such as lemmatization, could help reduce variability in the data by grouping words with similar meanings (e.g., "run" and "running") under a single root form, thereby improving topic coherence. Employing context-aware stop-word filtering to retain meaningful short words, such as “AI” or “God,” could preserve important contextual information that might otherwise be lost. By refining these preprocessing steps and increasing the number of topics in the LDA model, it would be possible to achieve a deeper, context-driven interpretation of the threads. This could enable the detection of subtler thematic nuances and improve the alignment of topics with user engagement and toxicity patterns.

Additionally, a more strategic sampling approach, such as matching the population characteristics rather than random selection, could yield results that better represent the broader Reddit ecosystem. It would also be valuable to analyze how Reddit discussions have evolved over time and how patterns have changed. While our dataset lacked temporal features, parsing the data differently—by leveraging the repository’s organization of posts by month—could enable such an analysis and provide deeper insights into the progression of Reddit’s discussion dynamics.

References

[1] Reddit. (n.d.). *Content Moderation, Enforcement, and Appeals*. Retrieved from Reddit Help.
[2] O'Donnell, L. (2024, March 7). *Reddit introduces AI-powered harassment filter*. Retrieved from The Register.
[3] CNA. (2024). *Reddit may need to ramp up spending on content moderation, analysts say*. Retrieved from ChannelNews Asia.
[4] Hugging Face. *Reddit Questions with Best Answers Dataset.* Hugging Face Datasets, n.d. Retrieved from https://huggingface.co/datasets/nreimers/reddit_question_best_answers
[5] Nudurupati, Sreeram. Essential PySpark for Scalable Data Analytics: A Beginner's Guide to Harnessing the Power and Ease of PySpark 3. Packt Publishing, 2021.
[6] Hugging Face. (n.d.). *unitary/toxic-bert*. Hugging Face. Retrieved from https://huggingface.co/unitary/toxic-bert