Toxicity in Reddit Discussion Threads: Impacts and Predictive Insights

Nahiyan Bin Noor[1], Niloofar Yousefi[1], Billy Spann[1], and Nitin Agarwal[1]

COSMOS Research Center,
University of Arkansas at Little Rock, AR 72204, USA
{nbnoor, nyousefi, bxspann, nxagarwal}@ualr.edu

Abstract. Harmful content, including abusive language, disrespect, and hate speech, is a growing concern on social media platforms. Despite efforts to tackle this issue, completely preventing the impact of harmful content on individuals and communities remains challenging. This paper utilizes Reddit data using a tree structure to study the impact of toxic content on communities. Machine learning algorithms are employed to classify the toxicity of each leaf node based on its parent, grandparent nodes, and the overall average toxicity of the tree. Our methodology aids policymakers in identifying early warning signs of toxicity and guiding harmful comments toward less toxic avenues. The research offers a comprehensive analysis of social media platform toxicity, facilitating a better understanding of variations across platforms and the impact of toxic content on different communities. Our findings offer valuable insights into the prevalence and impact of toxic content on social media platforms, and our approach can be used in future studies to provide a more nuanced understanding of this complex issue.

Keywords: Toxicity Visualization Tree · Toxicity Prediction Model · Reddit · Machine Learning.

1 Introduction

Social media platforms have experienced significant growth in recent years, enabling individuals to communicate efficiently and easily across the world. However, social media platforms have also witnessed an alarming increase in harmful content, such as hate speech, abusive language, and toxic comments. Social media toxicity is defined as the use of rude or disrespectful language, including threats, insults, and identity-based hate. It also encompasses harassment and socially disruptive persuasion like misinformation, radicalization, and gender-based violence [1, 2]. The increase of toxic content on social media platforms represents a significant threat to both individuals and communities in general. In this study, we focus on Reddit, one of the most popular social media platforms, to investigate the prevalence and impact of toxic content on this platform. Even though Reddit policymakers have imposed some rules and regulations that remove harmful content to prevent toxicity from their social media, yet they are still not able

to remove all toxic posts, specifically the root of toxic comments. NB. Noor [3] showed a comment thread in Reddit where Reddit authority removed one toxic comment; however, they could not detect the previous toxic comment and the root of that comment. Thus our research questions revolve around three main themes, which are the following: RQ1: How often do Reddit conversations end with toxic comments? RQ2: Are toxic Reddit conversations characterized by wider, deeper, and larger branches compared to non-toxic conversations? RQ3: Can we predict the toxicity of the final comment in a Reddit conversation by looking at the toxicity of the previous two comments and the overall toxicity of the conversation? To address these research questions, we employ an approach utilizing a tree structure to visually and comprehensively examine the impact of toxic content on Reddit communities. Our methodology enables policymakers to detect early warning signs of toxicity and redirect potentially harmful comments in less toxic directions, thereby mitigating the impact of toxic content on Reddit communities. Our study provides a comprehensive analysis of toxicity on social media platforms and allows for a deeper exploration of the impact of toxic content on individual communities. The findings of our study offer valuable insights into the prevalence and impact of toxic content on social media platforms, and our approach can be used in future studies to provide a more nuanced understanding of this complex issue.

2 Literature Review

Several studies have so far been done to classify toxic and non-toxic comments. To distinguish between toxic comments and non-toxic comments, Sahana et al. [4] and Taleb et al. [5] devised methods for binary categorization of toxicity. In another study, Kumar et al. [6] recommended a variety of machine-learning techniques to categorize toxic comments. Noor et al. [7] and DiCicco et al. [8] aim to detect toxicity and binary classify them, to evaluate the level of toxicity present in discussions related to COVID-19 different social media platforms. A study by Saveski et al. [9] conducted a study on Twitter comments and found that toxic conversations had larger and more complex conversation trees compared to non-toxic ones. They developed a method to predict the toxicity of the next reply in a conversation with up to 70.5% accuracy, providing insights into conversation toxicity and the likelihood of a user posting a toxic reply. Coletto et al. [10] addressed the challenge of identifying controversial topics in social media by using network motifs and local patterns of user interaction. Their approach achieved 85% accuracy in predicting controversy, outperforming other structural, propagation-based, and temporal network features by 7%. In another study, Backstrom et al. [11] highlighted the importance of algorithmic curation for online conversations, specifically focusing on discussion threads. They proposed and evaluated different approaches leveraging network structure and participant behavior, demonstrating that learning-based methods that utilize this information enhance performance in predicting thread length and user re-entry. Hessel et al. [12] concentrated on predicting controversial posts in

Reddit communities by examining textual content and tree structure of initial comments. They found that incorporating discussion features enhances predictive accuracy, particularly with limited comments, and noted that conversation structure features tend to have better generalization across communities than conversation content features. Additionally, Rajadesingan et al. [13] investigated the maintenance of stable norms in political subreddits on Reddit. They identified self-selection, pre-entry learning, post-entry learning, and retention as key processes, with pre-entry learning being the main factor contributing to norm stability. Newcomers adjusted their behavior to match the toxicity level of the subreddit they joined, but these adjustments were specific to the community and did not lead to transformative changes across political subreddits. Zhang et al. [14] introduce a computational framework for analyzing public discussions on Facebook. It captures social patterns and predicts discussion trajectories, highlighting participant tendencies over content triggers.

3 Methodology

3.1 Data Collection

In this study, the data was collected from the Reddit platform, specifically from the "r/Nonewnormal" subreddit, which is known for its anti-vaccine stance. The Pushshift API was utilized to retrieve all posts and comments from a database where the information is stored. The data collection process was implemented using the Python PSAW library, as described in its documentation. The dataset consisted of 2.2 million posts and comments collected from June 2020 to August 2021. The data for conducting this study were obtained from a subreddit that was banned in September 2021 due to its dissemination of misinformation and toxic content related to masks and vaccines [15, 16].

3.2 Data Cleaning and Pre-processing

The collected datasets underwent a cleaning process to eliminate non-intended language. Afterward, the posts and comments datasets were merged, excluding removed or deleted posts and comments along with their respective columns. Additionally, columns containing spam or bot-generated comments were removed. For this study, posts or comments without parent or child nodes were excluded. These steps were taken to guarantee that the datasets used for subsequent analyses were of high quality and dependable. As a result of these measures, the dataset was refined, and it contained 1.2 million entries after the cleaning and preprocessing steps.

3.3 Toxicity detection

Detoxify [17], developed by Unitary A.I., is a machine-learning model utilized in this research. It employs a Convolutional Neural Network (CNN) architecture trained with word vectors to classify text as toxic or non-toxic. The Detoxify API,

accessible to developers and researchers, provides a probability score between 0 and 1 for each input text, indicating the likelihood of being toxic, with closer to 0 being more nontoxic and closer to 1 being more toxic. The cleaned dataset was analyzed using the Detoxify model, generating toxicity scores across seven categories: Toxicity, Severe Toxicity, Obscene, Threat, Insult, Identity Attack, and Sexually Explicit. These categories provide insights into different levels and types of toxic content in online discussions. A threshold of 0.5 was employed to determine which texts were considered toxic, following a similar approach by Saveski et al. [9] who determined a threshold of 0.53 through annotator feedback. Texts with a toxicity score higher than 0.5 were labeled as toxic, while those with a score of 0.5 or lower were considered non-toxic. By utilizing the Detoxify model to analyze the dataset, the analysis provided valuable insights into the prevalence and characteristics of toxic text in online discussions. It is worth noting that previous studies have also utilized Detoxify and similar tools to calculate toxicity scores for online discussions. Comparing these scores across different research can help identify trends and patterns in the use of toxic language online, thereby informing the development of strategies and tools aimed at fostering more civil and respectful interactions on the internet.

3.4 Conversation tree generation

This study focused on the "Nonewnormal" subreddit and collected 23,000 conversation trees from this specific subreddit. To investigate the range of toxicity levels in the dataset, the average toxicity score was calculated for each conversation tree, and they were divided into five categories. A subset of 100 trees was selected from each category for further analysis, resulting in a total of 500 trees. The distribution of these trees across the five categories is illustrated in Figure 1: non-toxic, less toxic, moderately toxic, toxic, and most commented. This approach allowed for a more detailed examination of toxic conversations, taking into account the patterns and dynamics observed across different levels of toxicity within the dataset. The selection of the fifth category was based on user engagement, as varying degrees of toxic comments are prevalent within the Reddit community and are often considered a common form of discourse. We ignored the fourth category of the nontoxic tree as all of the comments and replies of this category were nontoxic. Figure. 1 depicts the clear distribution of all tree categories.

3.5 Conversation tree visualization

To visualize the conversation trees from Reddit, we utilized the Tulip visualization tool. Toxicity was determined based on comments with a toxicity score exceeding 0.5. In the visualization, red and blue colors were used to distinguish toxic and non-toxic comments, respectively. An example of a conversation tree from category 5, which had the highest comment count, is displayed in Figure 2. This particular tree comprised 528 comments, with 109 of them classified as toxic. It ranked 54th among the top 100 trees in that category based on its size.

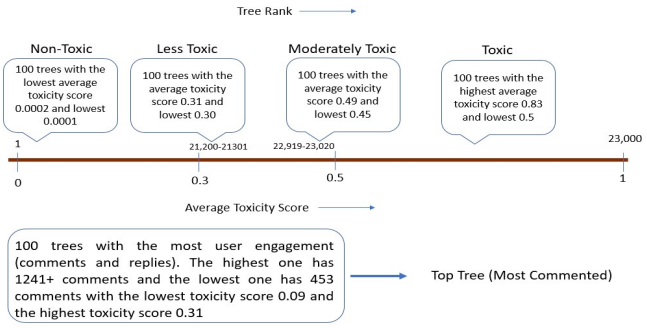

Fig. 1. Conversation tree generation and category distribution.

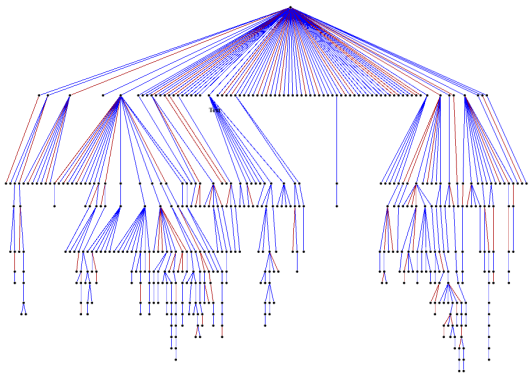

Fig. 2. A conversation tree visualization from most commented tree category (54th)

4 Results and Findings

This section presents the results and findings of this paper by answering the research questions mentioned in the introduction section.

4.1 Conversations End with Toxicity in Different Categories of Tree

The study's findings reveal that toxic conversations in online communities tend to diminish over time due to their inherently toxic nature. While these conversations initially attract participants, they eventually discourage further engagement. This pattern is particularly noticeable in communities with a significant presence of toxic behavior, such as the anti-vaccine community. To explore this phenomenon, the study analyzed conversation trees, and the results are presented in Figure 3, with blue bar graphs indicating nontoxic trees and orange

bar graphs indicating toxic trees. The analysis revealed that more than 50% of toxic conversations concluded with toxicity without any further response or engagement (Figure 3 - second plot). In moderate toxic conversations (Figure 3 - third plot), 42% still ended with toxicity, while in less toxic conversations (Figure 3 - fourth plot), 27.4% of toxic posts were still concluded by the user. Conversely, the conversation trees with the highest number of comments (Figure 3 - first plot) predominantly exhibited non-toxic characteristics. The toxicity scores in these conversations ranged from 0.09 to 0.31, indicating that they were primarily focused on general discussions. In the non-toxic conversations category, there were no instances of toxic comments. These posts mainly consisted of inquiries or sharing personal experiences. These findings suggest that toxicity has a negative impact on the sustainability of conversations in online communities. As toxicity increases, users are more inclined to disengage, leading to premature termination of discussions. Therefore, community managers and moderators should be attentive to toxic behavior and implement appropriate measures to mitigate it. This approach is vital for fostering healthy and sustainable online conversations.

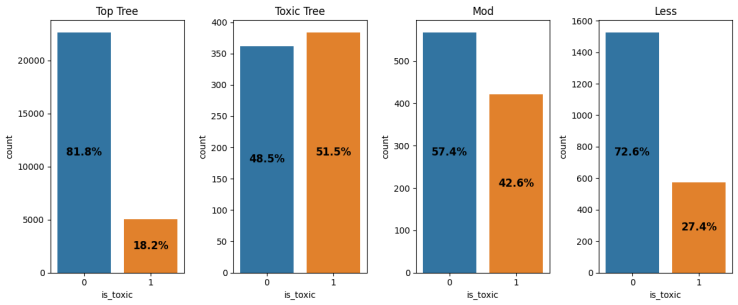

Fig. 3. Conversations that end with toxicity for each tree category

4.2 Structure of Toxic Conversations: Wider, Deeper, and Larger Branches

In this study, an analysis was conducted on a conversation tree belonging to the most commented tree category. This particular conversation tree was ranked 76 based on the number of comments and consisted of 495 comments, with 105 comments identified as toxic (highlighted by red edges in Figure 4). The average toxicity score for this conversation tree was determined to be 0.21. Visual analysis of the conversation tree revealed that toxic conversations tend to exhibit wider, larger, and deeper branches. The red edges are shown with black arrows in Figure 4 demonstrating that each toxic comment tends to generate more toxic comments, resulting in an expansion of the conversation both in width and depth (circled in red in Figure 4). Each branch within the conversation tree typically contains more nodes than the previous one, contributing to the overall size of the graph. Users tend to engage more with comments that receive numerous

replies, which further contributes to the development of toxic conversations. This

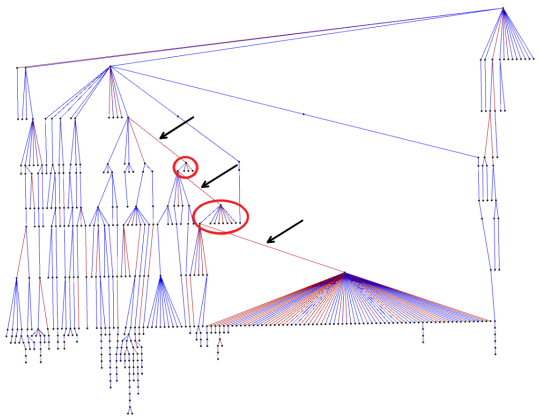

Fig. 4. Class distribution and accuracy of the prediction with different machine learning models for toxic tree

study offers visual representations of different conversation trees categorized by their toxicity levels and other attributes in order to address the research question of whether toxic Reddit conversations are characterized by wider, deeper, and larger trees. The selection of each tree was determined by specific criteria, such as comment quantity or toxicity level. Structural characteristics like the size, depth, and width of the trees and the occurrence of toxic comments within them were thoroughly examined. By using these visualizations, a clear depiction illustrates the progression of toxic conversations and their distinctions from non-toxic conversations. These findings can guide future research on online conversations and provide insights on how to encourage more positive and constructive interactions.

4.3 Predicting the Toxicity of the Last Message in a Reddit Conversation

In this section, a predictive analysis is conducted to predict the toxicity of the last reply in a conversation. This prediction was based on the toxicity scores of the previous two replies and on the overall toxicity score of the conversation. To perform the analysis, 100 trees from all categories were selected, except the non-toxic category. Specifically, all leaf nodes with two initial parents were extracted. Figure 5 illustrates the class distribution of all four tree categories.

To assess the effectiveness of various machine learning algorithms for the prediction task, such as Logistic Regression, Support Vector Machine, etc., the evaluation matrices for each category are shown in Table 1 with the algorithms

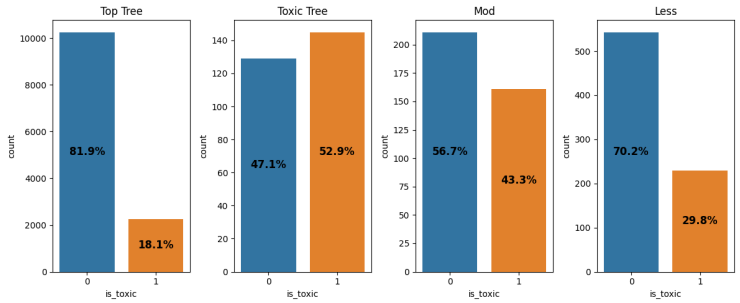

Fig. 5. Class distribution of all tree categories. First (most commented), second (most toxic), third (moderate toxic), and fourth (less toxic) plots

Table 1. Result of Toxicity Prediction for Each Tree Category

Tree Category	Accuracy	Precision	Recall	F1	Algorithm
Most Commented	0.82	0.86	0.82	0.75	Logistic Regression
Most Toxic	0.73	0.74	0.73	0.75	Adaboost Classifier
Moderate Toxic	0.59	0.59	0.59	0.59	Gradient Boosting
Less Toxic	0.70	0.66	0.70	0.65	Adaboost Classifier
Non Toxic	None	None	None	None	None

that are used to get the best accuracy. In Figure 5, we can see the best class distribution for toxic tree and for this category, the accuracy is 73% with the Adaboost classifier. Even though we obtained the highest accuracy for the most commented tree, there is unbalanced class distribution. In Figure 6 the confusion matrix is shown for each category of trees.

The findings indicate that it is possible to predict the toxicity level of the last reply in a conversation with a reasonable level of accuracy by taking into account the toxicity scores of the preceding two replies, and the overall toxicity score of the conversation. Once we can detect the toxicity in a comment thread, policymakers can end that conversation to protect potential spread of toxicity.

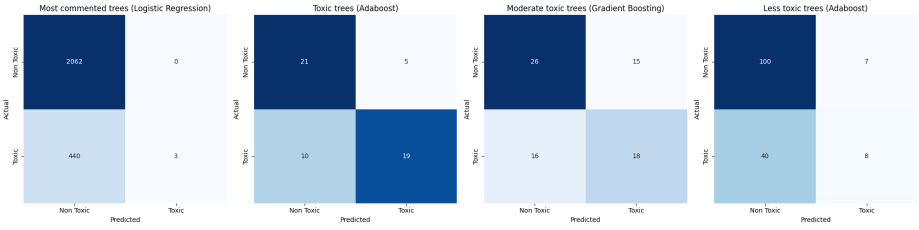

Fig. 6. Confusion matrix for the leaf node prediction from 100 toxic trees

5 Conclusion and Future Works

The prevalence and impact of harmful content on Reddit, such as abusive language and hate speech, is a concerning issue. This paper addresses the challenges in detecting and removing toxic comments and presents a tree structure approach to aid policymakers in the early detection and redirection of harmful content, mitigating its impact on Reddit communities.

Our observations indicate that toxic conversations often end with harmful remarks, leading to decreased engagement from users. However, these discussions tend to generate broader, more profound, and more extensive branches, suggesting that toxic interactions can have a long-lasting impact and influence on the behavior of other users. Moreover, our analysis shows that it is possible to predict the toxicity of the next response by considering the toxicity scores of the previous two comments and the overall context. This discovery carries significant implications for the advancement of moderation tools that can aid online platforms in effectively identifying and preventing toxic interactions.

In our future work, we will consider the class imbalance of toxic and nontoxic comments and try to imply different techniques to balance the dataset before applying machine learning algorithms. Moreover, we will also take into consideration what might be the toxicity of any following replies based on all previous replies. We also will examine which comments impact upcoming replies in a conversation thread, so that they can be diminished from social media for making safer places for all the users.

6 Acknowledgement

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920, IIS-1636933, ACI-1429160, and IIS-1110868), U.S. Office of the Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014-16-1-2016, N00014-16-1-2412, N00014-17-1-2675, N00014-17-1-2605, N68335-19-C-0359, N00014-19-1-2336, N68335-20-C-0540, N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Laboratory, U.S. Army Research Office (W911NF-20-1-0262, W911NF-16-1-0189, W911NF-23-1-0011), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059), Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock, and the Australian Department of Defense Strategic Policy Grants Program (SPGP) (award number: 2020-106-094). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

References

1. Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., and Leskovec, J.: Any- one Can Become a Troll: Causes of Trolling Behavior in Online Discus- sions. In: Proc.

ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW17), ACM Press, pp. 1217, (2017) https://doi.org/10.1145/2998181.2998213
2. Yousefi, N., Noor, N. B., Spann, B., Agarwal, N.: Towards Developing a Measure to Access Contagiousness of Toxic Tweets. In: TrueHealth 2023, Workshop on Combating Health Misinformation for Social Wellbeing, In press, (2023).
3. Noor, N. B.: Toxicity and Redditv: A study of harmful effects on user engagement and community health (Order No. 30423680). Available from Dissertations & Theses @ the University of Arkansas at Little Rock, (2806341066), (2023).
4. Sahana, B. S., Sandhya, G., Tanuja, R. S., Sushma Ellur, and Ajina, A.: Towards a safer conversation space: detection of toxic content in social media (student consortium). In: IEEE Sixth International Conference on Multimedia Big Data (BigMM), pp. 297-301. IEEE, (2020).
5. Taleb, M., Hamza, A., Zouitni, M., Burmani, N., Lafkiar, S., and En-Nahnahi, N.: Detection of toxicity in social media based on Natural Language Processing methods. In: International Conference on Intelligent Systems and Computer Vision (ISCV), pp. 1-7. IEEE, (2022).
6. Kumar, A.K., and Kanisha, B.: Analysis of multiple toxicities using ML algorithms to detect toxic comments. In: 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp. 1561-1566. IEEE, (2022).
7. Noor, N. B., Yousefi, N., Spann, B., and Agarwal, N.: Comparing Toxicity Across Social Media Platforms for COVID-19 Discourse. In: The Ninth International Conference on Human and Social Analytics, (2023).
8. DiCicco, k., Noor, N. B., Yousefi, N., Spann, B., Maleki, M., and Agarwal, N.: Toxicity and networks of COVID-19 discourse communities: A tale of two media platforms. In: The 3rd Workshop on Reducing Online Misinformation through Credible Information Retrieval, (2023).
9. Saveski, M,. Roy, B., & Roy, D.: The structure of toxic conversations on Twitter. In: Proceedings of the Web Conference 2021, pp. 1086-1097, (2021).
10. Coletto, M., Garimella, K., Gionis, A., Lucchese, C.: Automatic controversy detection in social media: A content-independent motif-based approach. In: Online Social Networks, and Media, Volumes 3–4, Pages 22-31, ISSN 2468-6964, (2017) https://doi.org/10.1016/j.osnem.2017.10.001
11. Backstrom, L., Kleinberg, J., Lee, L., & Danescu-Niculescu-Mizil, C.: Characterizing and curating conversation threads: expansion, focus, volume, re-entry. In: Proceedings of the sixth ACM international conference on Web search and data mining, pp. 13-22, (2013).
12. Hessel, J., & Lee, L.: Something's brewing! Early prediction of controversy-causing posts from discussion features. In: arXiv preprint arXiv:1904.07372, (2019).
13. Rajadesingan, A., Resnick, P., & Budak, C.: Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 14, pp. 557-568, (2020).
14. Zhang, J., Danescu-Niculescu-Mizil, C., Sauper, C., & Taylor, S.J.: Characterizing online public discussions through patterns of participant interactions. In: Proceedings of the ACM on Human-Computer Interaction, 2(CSCW), 1-27, (2018)
15. https://arstechnica.com/tech-policy/2021/09/reddit-bans-r-nonewnormal-and-quarantines-54-covid-denial-subreddits/
16. https://www.forbes.com/sites/carlieporterfield/2021/09/01/reddit-bans-controversial-covid-subreddit-after-users-protest-disinformation/?sh=16870c905a2a
17. https://github.com/unitaryai/detoxify