DNDS6014 Introduction to Computational Social Science
Project Report Fall 2021-22

Toxicity at Scale?

A Structural View of Conversation quality in political subreddits

Yijing Chen

chen_yijing@phd.ceu.edu

December 13, 2021

Abstract

abstract here

Keywords
Social media, network analysis, political communication, friendship paradox, toxic behaviors

1 Introduction

Deliberative democracy asks for efficient political communication that meaningfully engages the public in discussions of key issues. Social networking sites (SNS) provide a shared cyberspace with flexible community boundaries and topical agendas that facilitates political communication. However, virtual discussions are always as meaningfully as expected, and sometimes even become brewing sites of toxic content.

To better understand and reflect upon the political deliberation processes that happen online, researchers need to customize analysis frameworks for different platforms with varying affordances of communication. In the field of network science, for instance, enormous amount of work have constructed who-follows-whom networks to capture the flow of information from followees to their followers [e.g., 1, 2, 3]. While this intuitive approach works reasonably well in characterizing users or communities on follower-following based platforms such as Twitter, it might not be the best and only option for analyzing forum-based platforms such as Reddit, where the information feed delivered to individual users are not fundamentally determined by whom they are following.

Hence, with a focus on the toxic content in political subreddits, this project is interested in exploring creative ways to construct and interpret network models applied in the context of Reddit, and hopes to answer the overarching question: where and how toxic conversations happen online.

More specifically, I observe toxic observations at two different levels. At the user level, I hope to answer:

1. Who talks toxically?
2. Toxicity Paradox (TP): are my neighbors more toxic than me?
3. Temporal Toxicity Paradox (TTP): throughout my Reddit lifespan, are my neighbors consistently more toxic than me?

Further at the submission level, I ask the following questions:

1. How well do toxic conversations engage users?
2. Where is the most toxic post usually located?

2 Dataset and Preprocessing

This project analyzes a longitudinal Reddit dataset from January 2012 to December 2015. Reddit is a hub website with many self-moderated subreddits (i.e., communities), which users can subscribe to and receive subreddit updates on their front page timelines. Within each subreddit, users can create submissions and leave comments under a submission or a comment.

I use Pushshift API[1] to collect user-generated content in 51 active political subreddits. The dataset contains over 1.4 million submissions and over 20 million comments created by more than 900 thousands unique users. Each entry

[1] Pushshift Reddit API documentation on GitHub: https://github.com/pushshift/api.

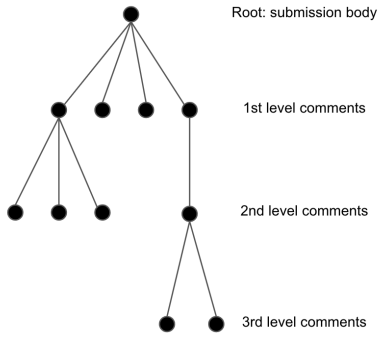

Figure 1: A toy example of a tree graph representation for a submission

is time-stamped with information about its author, textbody, net upvote score and toxicity score[2].

Given limited time and computational resources, I truncate the data into yearly units for analysis, and will only present the result of 2012 subset in this report, if no significant variation is noticed across years. I also filter out inactive users who have less than 100 posts (i.e., either submissions or comments) or have appeared in less than 2 unique subreddits and 10 unique submissions. For user-level analysis, I also remove submissions and comments displaying `[deleted]` in the author column or having an empty textbody, all of which are kept in submission-level analysis so that the tree graph construction can preserve the original structure of conversations as much as possible.

It is worth mentioning that, despite of the publicity of the dataset, users would not normally create the content for research purpose in particular, nor would them expect their digital traces to be made publicly identifiable. To minimize privacy risks, I only analyze and discuss the dataset in the academic setting, and make sure that personally identifiable information is either aggregated or anonymized in final results. Raw data or text content will not be released in the project repository.

3 Methods

3.1 User-level analysis

3.1.1 User-to-user network

I construct an undirected user-to-user graph $G_u = \{V_u, E_u\}$ where each node represents an individual user. I connect two users if they have co-appeared in the same submission at least once, assuming that they have participated in a shared conversation and have some forms of interactions, either directly through replies or indirectly through content browsing. Starting from the most basic question–who talks toxically, I use the user graph to characterize users' activity levels by node degrees, and examine the correlations between degree d_i and maximum/average toxicity levels. By looking at the total number of unique neighbors a given user has shared conversations with, this measurement of activity levels positively correlates with the raw counts of submissions and comments, yet offers more nuance information about the scope of user navigation by accounting for the number of participants one have encountered along the way.

[2]Toxicity scores are assigned by Perspective API: https://www.perspectiveapi.com/.

	Feature
Text	total string length
	average string length
Tree	# of unique users
	# of branches
	maximum depth
	structural virality

Table 1: Measures of conversational engagement

In addition to computing correlations for the empirical network, I generate two shuffled networks as null models for comparison–the first shuffling toxicity per-post and the second shuffling toxicity per-user, so that I can disentangle the effect of of varying user nature from that of pure aggregation. In other words, after empirically observing how toxicity scores distribute across user groups with varying activity levels, I wonder whether the distribution is mainly a result of the actual variation in users' toxicity levels with different navigation scopes (e.g., active users are by nature more toxic), or is simply caused by the aggregating process (i.e., active users with more posts are more likely to appear more toxic).

3.1.2 Toxicity paradox: aggregated and temporal snapshots

After individually correlating users' toxicity and activity levels, I proceed to analyze users' toxicity levels within their local neighborhood. By comparing their average toxicity scores with those of their neighbors, I investigate whether the Toxicity Paradox (TP, i.e., my neighbors are on average more toxic than me) holds true, both per node and per network. For each node, TP is true if its average toxicity x_i is lower than the average toxicity of its neighbors $\mathcal{N}(i)$ [5], namely:

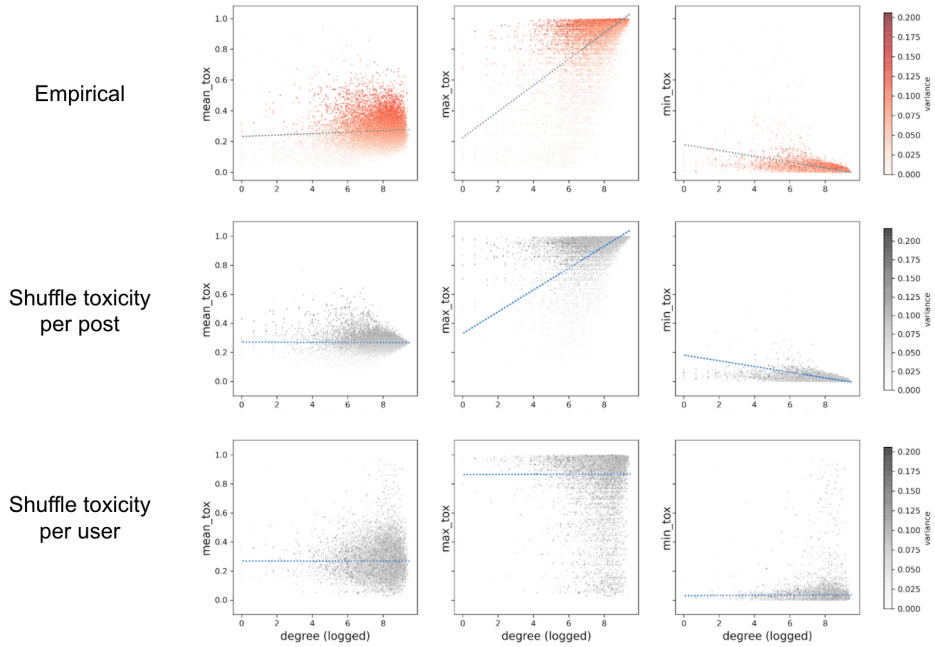

Figure 2: Correlation analysis of user degree and toxicity level for the empirical network and two shuffled networks. k is the slope of the OLS fitting line.

Figure 3: Illustrations of broadcast (left) and viral diffusion (right) [4]

$$x_i < \frac{\sum_{j \in \mathcal{N}(i)} x_j}{k_i}$$

For the entire network, TP is true if the expected level of toxicity across all nodes is lower than that across all neighbors [5], that is:

$$\langle x \rangle < \langle x \rangle_{nn} = \frac{\sum_{i=1}^{N} k_i x_i}{\sum_{i=1}^{N} k_i}$$

To stretch one step further from Eom and Jo's work on generalized friendship paradox, I examine node-level TP temporally throughout users' posting lifespan (i.e., Temporal Toxic Paradox, TTP). To operationalize, I retrieve all their posts (i.e., submissions and comments) for a given user i, and define the lifespan as the period of time starting from their first post and ending at their last. Then, I create a sliding window with the length of 10% times their entire lifespan. Within each window, I identify all temporal neighbors $\mathcal{N}(i)^{(t)}$ of user i, and calculate the temporal average toxicity scores for the user and all neighbors. Similarly, for TTP to be true per node within window t:

$$x_i^{(t)} < \frac{\sum_{j \in \mathcal{N}(i)^{(t)}} x_j}{k_i^{(t)}}$$

3.2 Submission-level analysis

3.2.1 Submission tree graph

For submission-level analysis, I focus on the relationship between conversational structures and toxicity levels, for which I build a tree graph to represent submission with at least one comment. As shown in Figure 1, each node is a single post, and a link exists if one post is a reply to another. The depth of node indicates its nested level in a submission.

3.2.2 Variations of toxicity across submissions

To measure the extent to which a submission engages a certain pool of participants, I use both textual and tree features to obtain submission-level metrics (see Table 1). For tree features in particular, apart from basic metrics such as the number of branches and maximum depth, I also consider the structural virality, a metric proposed by Goel, Anderson et al. [4] to account

Network	Independent var.	OLS slope	Pearson's r
empirical	max toxicity	0.0852**	0.4498**
	mean toxicity	0.0047**	0.0658**
	min toxicity	-0.0191**	-0.2787**
shuffled per-post	max toxicity	0.0785**	0.4168**
	mean toxicity	0.0011**	0.0029**
	min toxicity	-0.0184**	-0.2857**

Table 2: OLS slope and Pearson's r for the empirical and per-post shuffled network (** indicates that the result is statistically significant with a p-value lower than 0.01).

Network	ρ_{kx}	r_{xx}	H	$\langle x \rangle$		$\langle x \rangle_{nn}$
Empirical	0.0658**	-0.0001	0.6170	0.2674	$<$	0.2724
Shuffle toxicity per post	-0.0035	-0.0001	0.6170	0.2672	$\approx$	0.2673

Table 3: Network-level toxicity paradox statistics. For both empirical and per-post shuffled network, I calculate the Pearson's r between toxicity and degree ρ_{kx}, the assortativity of node toxicity r_{xx}, the average paradox holding probability H, the average toxicity of nodes $\langle x \rangle$ and of neighbors $\langle x \rangle_{nn}$ (** indicates that the result is statistically significant with a p-value lower than 0.01).

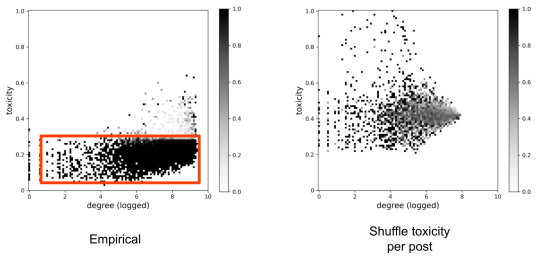

Figure 4: The paradox holding probability as a function of degree and average toxicity scores.

for both the span and the depth by calculating the average distance between all node pairs in a tree T, namely:

$$v(T) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}$$

Originally applied in analyzing diffusion trees in Twitter network, structural virality provides an interpolating measure between two extreme cases of information diffusion processes–single-generation broadcast and multi-generation spreading (see Figure 3).

3.2.3 Variations of toxicity within a submission

Beyond per-tree analysis, I locate the most toxic post and see how early it occurs within a submission. Additionally, I compare the toxicity levels of parent nodes and their corresponding child nodes to examine whether there exists a generalized pattern of toxicity progression, that is, whether users tend to receive more toxic replies than their own posts.

4 Results and Discussions

4.1 User-level

4.1.1 Who talks toxically?

To begin with, I create Figure 2 to demonstrate how average, maximum, and minimum toxicity levels correlate with users' degree in both the empirical and shuffled networks. If only looking at the empirical sub-figure (top row), we may conclude that the highest toxicity levels achievable for a given user positively correlates with the number of neighbors, while average and minimum toxicity are slightly correlated but almost invariant. However, this does not necessarily imply that the more active a given user is, the more toxic the user can potentially be, and I will elaborate on breaking down the positive correlation by comparing the result against null models.

Comparing the result between the empirical the per-user shuffling network (bottom row), we can see that randomly assigning toxicity scores to individual users would flatten out all OLS fitting lines as expected, which, however, does not serve as a meaningful null model as the post-aggregation shuffling cannot inform us of possible biases that arise from the aggregation process itself. Hence, I argue that per-post shuffling that disarranges toxicity at the post level before by-user aggregation produces a better null model to compare against. Thus, I focus on comparisons between the empirical and the per-post shuffled networks and report the OLS fitting slope and Pearson's r in Table 2. In line with the visual impression from Figure 2, two networks show fairly similar distributions of toxicity levels with respect to activity levels, with numerically close Pearson's correlation coefficients

and OLS slopes. This implies that the correlation between toxicity and node degree is largely attributable to the by-user aggregation, meaning that the more posts you create, the more likely that you will have a wider range of toxicity scores with a higher ceiling and a lower floor, even when the toxicity of your posts are randomly sampled from a given pool of toxicity scores.

4.1.2 Toxicity paradox

Next, I look at toxicity paradox in yearly aggregates. For node-level TP, I produce Figure 4 to show the paradox holding probability $h(k, x)$ with respect to node degree k and average toxicity x. Unlike the null result, the empirical result shows a dense TP region at the right bottom corner of the plot, implying that TP is generally true for nodes with average toxicity below 0.3 and logged degree over 4. Interestingly, we can observe a sharp horizontal division at around $x = 0.3$: the paradox holding probability would drop drastically if we slide from users with toxicity levels slightly above 0.3 to those below 0.3. Such an abrupt rather than gradual decrease in $h(k, x)$ brings up an intriguing direction for future explorations.

For network-level TP, I calculate the Pearson's r between toxicity and degree ρ_{kx}, the assortativity of node toxicity r_{xx}, the average paradox holding probability H, and the average toxicity of nodes $\langle x \rangle$ and of neighbors $\langle x \rangle_{nn}$. Results from the empirical and per-post shuffled networks are reported in Table 3. Overall, neither of the network have an impressively high paradox holding probability, which, in accordance with Eom and Jo's finding [5], can be explained by the fact that the correlation between degree and average toxicity is very weak, and that the toxicity assortativity is close to zero. Comparing results from two networks, I notice that with a weaker and non-significant toxicity-degree correlation, the net difference between $\langle x \rangle$ and $\langle x \rangle_{nn}$ in null model is even smaller, which further confirms the insights from Eom and Jo's work that one origin of paradox roots in the positive correlation between degree and a given node characteristic.

In addition to aggregated analysis, I break down TP in temporal units throughout users' lifespan and show results in Figure **??**

4.2 Submission-level

5 Conclusions

Acknowledgements

The dataset was collected by the author in May 2020; the author then performed parallelized toxicity scoring jobs on University of Michigan's Great Lakes Slurm cluster, which was funded by Professor Abigail Jacobs during her work as a research assistant in summer 2020.

References

[1] Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *Fifth international AAAI conference on weblogs and social media*, 2011.

[2] Jiang Yang and Scott Counts. Predicting the speed, scale, and range of information diffusion in twitter. In *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[3] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74, 2011.

[4] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. The structural virality of online diffusion. *Management Science*, 62(1):180–196, 2016.

[5] Young-Ho Eom and Hang-Hyun Jo. Generalized friendship paradox in complex networks: The case of scientific collaboration. *Scientific reports*, 4(1):1–6, 2014.