Diffusion of Toxicity on Reddit discussions

Basiakou Kalliopi, Kotsinas Georgios

Online Social Networks and Media, dept. of Computer Science and Engineering University of Ioannina

ABSTRACT

Aggregation on social media has been investigated in a variety of scenarios, social media platforms, and modeled on different data using machine and deep learning to enable automatic identification and control of this activity. Because of the increased toxicity and violence in their own online community, users may be persuaded to act aggressively or even bully others. As a result, this conduct can propagate from one user to another and therefore, spread across online discussions. In this paper, we approach the crucial issue of social media aggression by examining and analyzing the discussions in different communities of reddit platform. To accomplish this, we build discussion trees for every submission collected, with root the initial post and children its comments and replies. Every text of the tree is replaced with a toxicity score, generated from Perspective API. Afterwards, we use these scores and compare them to different attributes of the whole trees, to see how toxicity correlates with the nature of the discussion. Finally, we examine the diffusion of toxicity by isolating the toxic nodes and studying their corresponding subtree.

1. Introduction

Online social media platforms provide unrivalled communication options, but they also bring with them unwelcome nasty conduct. Cyberbullying, racism, hate speech, and prejudice are some of the online aggressive behaviours that appear on such platforms, and they frequently have severe implications for individual users as well as for the society. Aggression can be apparent by uploading inappropriate content, such as bad sentiments and humiliating images, or implicit by inadvertently harming other online users, such as by spreading negative rumours. [1] Because of the increased toxicity and violence in their own social circle, users might be tempted to act violently and even harass others, as illustrated in [3]. Similar manifestations of this behaviour exist online, where anger can spread from one person to another. In reality, early sociological and psychological studies put out computer abuse models based on social learning, social ties, and planned behaviour theories [4]. To this purpose, [5] earlier research looked at how bullies can affect others in an online social network as well as the pairwise exchanges between users. This paper takes the first, but crucial steps to investigate toxicity propagation to Reddit discussions, and tries to answer the questions: *Does the size of discussions relates with the toxicity of the original post? How far the toxicity diffuses in the discussion? Does the votes of the initial post relates to the average toxicity of the discussion?*

Following the introduction, the paper is structured as follows: Section 2 discusses the methodology we followed; then, Section 3 presents the experimental analysis and lastly, Section 4 concludes our work.

2. Methodology

2.1 Data Collection

The first step is the data collection from Reddit API. Reddit is an American social news aggregation, content rating, and discussion website. Registered users submit content (submissions) to the site such as links, text posts, images, and videos, which are then voted up or down by other members. Posts are organized by subject into user-created boards called "communities" or "subreddits". Submissions with more upvotes appear towards the top of their subreddit and, if they receive enough upvotes, ultimately on the site's front page. Reddit administrators moderate the communities. Moderation is also conducted by community-specific moderators, who are not Reddit employees. Moreover, reddit's submissions can be sorted by different kind of characteristics, the most common are: "hot" sorting comes up with posts that's been getting a lot up upvotes/comments recently, "top" sorting for posts that has gotten the most upvotes over the set period, "controversial" sorting for posts that're getting multiple downvotes and upvotes and many other types of sorting. We collected our data from three different subreddits: r/politis, r/sports/, r/ukraine. For each sub-reddit we extracted posts from hot, top and controversial sorting along-side with their comments. We collect a post if it has minimum 20 and maximum 4000 comments, so that we can perform an analysis. For the data collection we used PRAW. To collect the data we need an authorized Reddit instance. For this purpose we needed a Client ID, a Client secret, a User agent, our Reddit username and the password. We provide these by passing in keyword arguments when calling the initializer of the Reddit class. To obtain a subreddit instance, we pass the subreddit's name when calling subreddit on our Reddit instance. Then we iterate through some of its submissions. We choose the sort we want to iterate through eg. controversial and the method will immediately return a ListingGenerator, which is to be iterate through. Submissions have a comments attribute that is a CommentForest instance. To iterate over all comments as a flattened list we call the list() method on a CommentForest instance.

2.2 Measure Toxicity

After data collection, every post and comment gets a toxicity score with values from 0 to 1 by the Perspective's (API) main attribute "TOXICITY". We define a comment as toxic if its Perspective score is higher than 0.45. Perspective API (https://www.perspectiveapi.com/) uses machine learning models to identify abusive comments. The models score a phrase based on the perceived impact the text may have in a conversation. Perspective models provide sores for several different attributes. In addition to the flagship TOXICITY attribute, Perspective can provide scores for attributes like SEVERE TOXICITY, INSULT, PROFANITY, IDENTITY ATTACK, THREAD, SEXUALLY EXPLICIT. The API is hosted on Google Cloud Platform, which means it can be used with any popular programming language. When you send a request to the API, you'll request the specific attributes you want to receive scores for. A comment is the text to be scored, could be a single post to a web page's comments section, a forum post, a message to a mailing list, a chat message, etc. The score is returned within the API response. The only score type currently offered is a probability score. It indicates how likely it is that a reader would perceive the comment provided in the request as containing the given attribute. For each attribute, the scores provided represent a probability, with a value between 0 and 1. A higher score indicates a greater likelihood that a reader would perceive the comment as containing the given attribute. For example, a comment like "You are an idiot" may receive a probability score of 0.8 for attribute TOXICITY, indicating that 8 out of 10 people would perceive that comment as toxic. Perspective API train each machine learning model on millions of comments from a variety of sources, including comments from online forums such as Wikipedia and The New York Times, across a range of languages. For each comment 3-10 raters who speak the relevant language annotate whether a comment contains an attribute (e.g., TOXICITY) following instructions below. Then post-process the

annotations to obtain labels by calculating the ratio of raters who tagged a comment as each attribute. As a result, if 3 out of 10 raters tagged a comment as toxic, train the API models to provide a score of 0.3 to this and similar comments. Raters are given a list of online comments. For each comment, their job is to read the comment, then if the comment is in a foreign language or not comprehensible for another reason (e.g. gibberish, different dialect, etc.), indicate that by selecting the checkbox, then choose the level of toxicity in the comment, selecting either “Very Toxic”, “Toxic”, “Maybe - I’m not sure” or “Not Toxic”. In the end, they answer a set of questions about the comment choosing from “Yes”, “Maybe - I’m not sure” or “No”. Example questions: “Does this comment contain identity-based hate?”, “Does this comment contain insulting language?”, “Does this comment contain threatening language?”. If in doubt, raters are asked to err on the side of “Yes” or “I'm not sure”. Raters have the opportunity to provide free-form additional details on their reasoning in tagging the comments.

2.3 Discussion trees construction and toxic nodes detection

Next step is the construction of the discussion trees. For every post, a tree was built with root the initial post and for children its comments. The depth of tree depends on the replies of each separated comment [example]. For every tree we save the following attributes: the root’s toxicity (Perspective score), the number of its comments (all nodes of the discussion tree), the average toxicity of the tree (the sum of toxicity scores of every tree node divided by the total number of tree nodes), levels of the tree (depth of the conversation based on the comments’ replies), the toxic ratio of the tree (the percentage of toxic comments out of all comments), the score of the votes (the up-votes minus down-votes of the post) and the upvote ratio (the percentage of the up-votes from all votes of the post.

Furthermore, as we traverse the trees we save the comments defined as toxic and analyze their corresponding sub-tree with the toxic node as root. For every sub-tree with root the toxic comment we collect: the toxicity of the root (Perspective score), the number of replies (including the replies of the replies), the average toxicity of the sub-tree (the sum of toxicity scores of every sub-tree node divided by the total number of sub-tree replies), levels of the sub-tree (depth of the conversation based on the replies), the score of the votes (the up-votes minus down-votes of the toxic comment) and the level of the first and the last toxic reply.

3. Experimental analysis

In this section of the paper we will present the experimental analysis on reddit data collected. Our aim is to study how toxicity correlates with the size of discussion and examine the spread of toxicity in the trees. The datasets used for the analysis are r/politics (top), r/sorts (controversial), r/ukraine (controversial) and r/ukraine (hot). The details of data are presented below in the table (Table 1). As we observe, in most of cases, toxicity scores remain low because reddit already uses Perspective for hosting healthier online conversations.

Table 1: Datasets Description

	r/politics top	r/sports controversial	r/ukraine controversial	r/ukraine hot
initial posts/trees	21	54	32	27
sum of comments	9801	8046	3634	5869
average root toxicity	0,14	0,15	0,15	0,13
average post's comments (first level children)	491	150	140,8	218,3
average toxicity of tree	0,15	0,17	0,2	0,18
average tree levels	10,9	9,8	9,4	9,6
average toxic ratio	0,23	0,25	0,27	0,23
average score of post	104553,4	356,7	61	3919,8
average upvote ratio	0,89	0,57	0,58	0,98

3.1 Analysis of trees

For the experimental analysis, we measured the correlation between the initial post's toxicity and: the average toxicity of the tree, the depth of the tree, the total votes of the post, the upvote score and the total comments/replies of the post. Furthermore, we measured the correlation between the average toxicity of the tree and: the depth of the tree, the total votes of the post, the upvote score and the total comments/replies of the post. Finally, we examined how the percentage of toxic comments is associated with the depth of the trees, the total votes and the upvotes ratio.

i. Measurement of post toxicity in relation to average toxicity of tree

In two out of four data sets we observe that as root toxicity increases the average toxicity of the tree also shows a slight increase

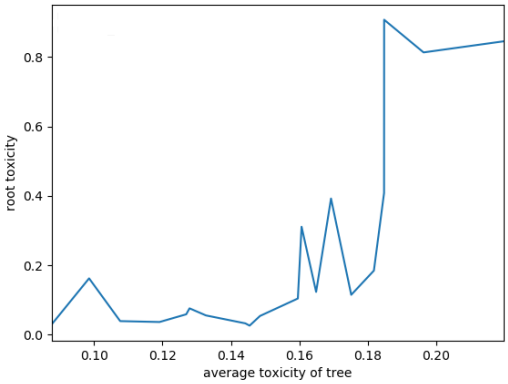

Figure 1. Dataset: r/political (top)

ii. Measurement of post toxicity in relation to levels of the tree

Comparing initial post's toxicity with tree depth (total levels) we do not observe any correlation between them, in any of the data sets.

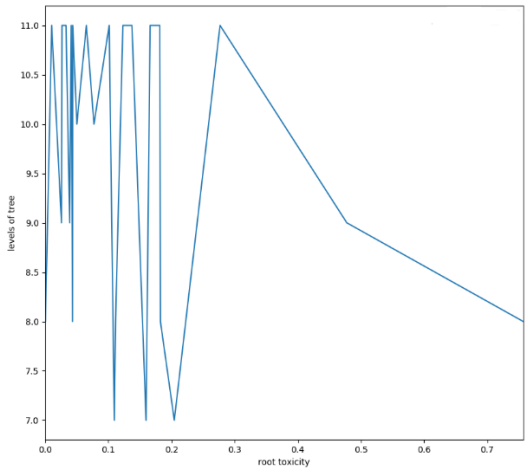

Figure 2. Dataset: r/ukraine (hot)

iii. Measurement of post toxicity in relation to score of the initial post

Comparing root's toxicity with the score of the post (upvotes-downvotes) we do not observe any correlation between them, in any of the data sets.

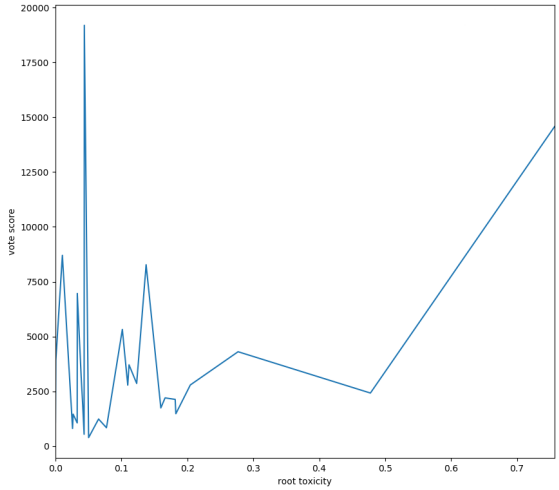

Figure 3. Dataset: r/ukraine (hot)

iv. Measurement of post toxicity in relation to upvote score

Comparing root's toxicity with the upvote score we do not observe any correlation between them, in any of the data sets.

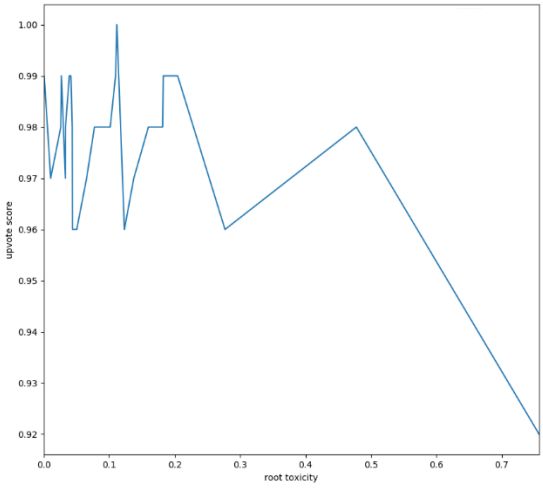

Figure 4. Dataset: r/sports (controversial)

v. Measurement of post toxicity in relation to num of comments

Comparing root's toxicity with the number of comments we do not observe any correlation between them, in any of the data sets.

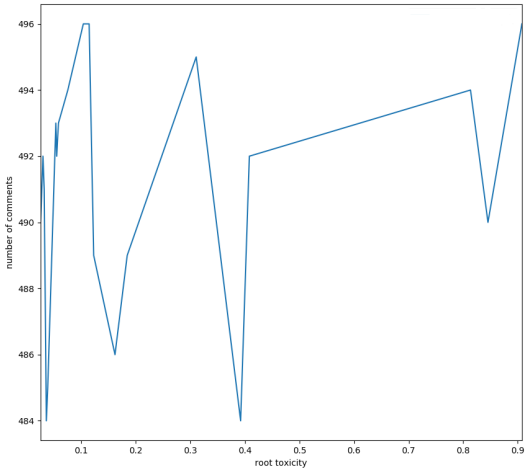

Figure 5. Dataset: r/sports (hot)

vi. Measurement of post toxicity in relation to toxic ratio

In two out of four data sets we observe that as root toxicity increases the toxic ratio of the tree also shows to be increase.

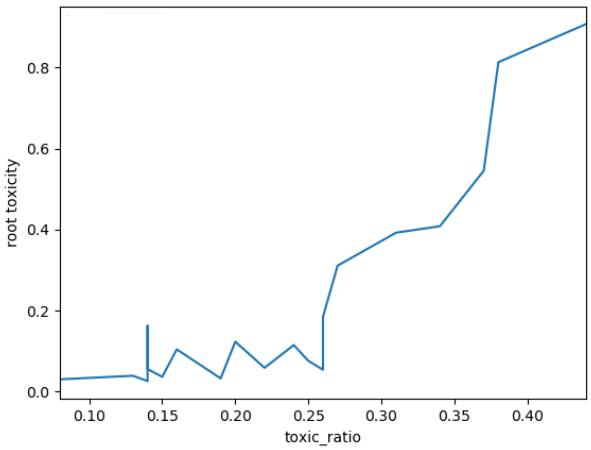

Figure 6. Dataset: r/politics (top)

vii. Comparing the average toxicity of the post with the depth of the tree, the number of comments and the score of the initial post no correlation detected.

viii. Measurement of average toxicity in relation to upvote ratio

One of the datasets shows that as the average toxicity increases the upvote ratio slightly increases as well.

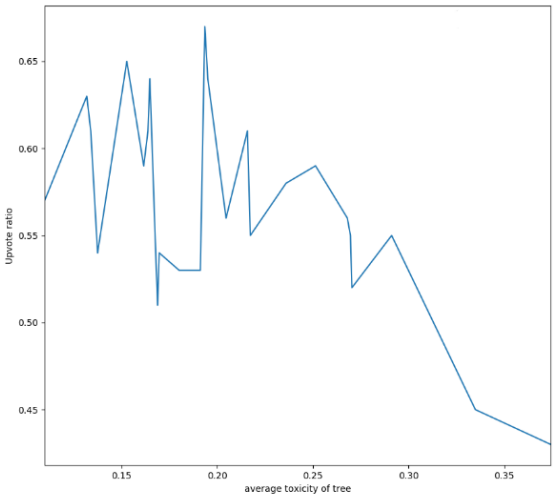

Figure 7. Dataset: r/ukraine (controversial)

3.2 Analysis of toxic comments

To examine the spread of toxicity in the tree, we made the following measurement. For each toxic node that we set as a root in its corresponding subtree, we stored at which level the first toxic response was found and at which the last one. By subtracting the depth of the last response minus the first we can see the propagation with respect to the total depth of the tree.The results for each dataset are presented in the table below (Table 2). We observe that in almost all cases after a very toxic comment there is no reply (~80% of subtrees). In the other cases, we mark that the spread of toxicity stops quite early (the spread does not exceed ~20% of the subtree)

Table 2: Toxicity Diffusion in Different Datasets

	r/politics top	r/sports controversial	r/ukraine controversial	r/ukraine hot
ndt_trees_ratio	81,70%	73,50%	76,30%	77,80%
diffusion_ratio	18,24%	18,40%	18,60%	19,43%

ndt_tree_ratio: Total subtrees with toxic node as root, that have no replies below, divided by the total subtrees. (Average percentage)

diffusion_ratio: Average percentage of the depth of the tree to which the toxicity has spread

For the experimental analysis, we measured the correlation between the toxic comment's toxicity and the average toxicity of the subtree, the depth of the subtree, the total votes of the toxic comment and the total replies of the toxic comment. Furthermore, we measure the correlation between the average toxicity of the toxic comment and the depth of the tree, the total votes of the comment and the total replies of the post. Finally, we examine the toxicity propagation by studying at which level of the tree we find the first toxic response and a where the last one occur.

i. Measurement of root toxicity of subtree in relation to average toxicity

As the toxicity of a comment increases the average toxicity of the subtree increases too. This happens because there are few (or none) replies, therefore the toxicity of the initial comment affects the average toxicity of its subtree, as we can see in the figure below.

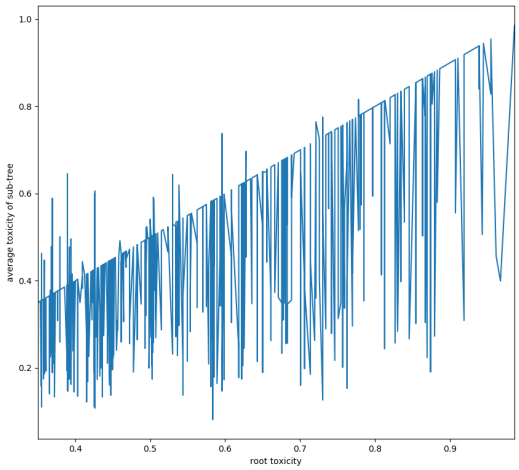

Figure 8. Dataset: r/ukraine (hot)

ii. Measurement of root toxicity of subtree in relation to score

Any correlation pointed out between subtree's root toxicity and its score (upvote-downvote)

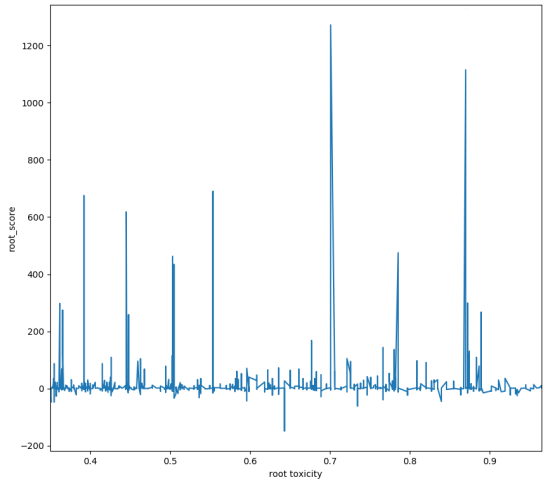

Figure 9. Dataset: r/ukraine (hot)

iii. Measurement of root toxicity of subtree in relation to the number of comments

Comparing the root toxicity of the subtree with the number of its comments no correlation detected.

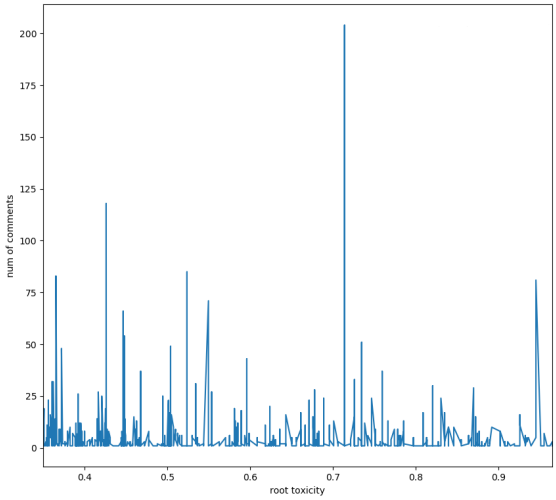

Figure 10. Dataset: r/ukraine (hot)

iv. Measurement of root toxicity of subtree in relation to the depth of the subtree.

As the toxicity of a comment increases the levels of the subtree decreases because there're only few (or none) after a toxic comment.

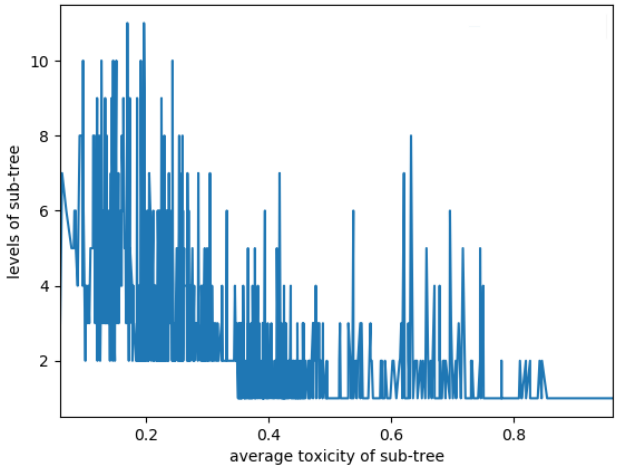

Figure 11. Dataset: r/politics (top)

4. Conclusion

In this paper real-time data from reddit analyzed. We focused on the study of toxicity and its spread in discussions, taking into account different characteristics of them, despite the fact that reddit filters out many abusive and insulting comments. The main feature around which the analysis was done, is the "Toxicity Score", obtained by analyzing the texts with the Perspective API. The conclusions we reached were that the toxicity of the original post is related to the average toxicity of the discussion, as well as to the percentage of toxic comments in the discussion. Furthermore, we found that the toxicity of the original post is not related to the size of the discussion and the votes a post has received. Finally, analyzing

individually the toxic comments found in discussions, we concluded that users rarely respond to toxic comments and when they do, the toxicity is not reproduced in the depth of the discussion.

5. References

[1]Marinos Poiitis, Athena Vakali, Nicolas Kourtellis: On the Aggression Diffusion Modeling and Minimization in Twitter. ACM Trans. Web 16(1): 5:1-5:24 (2022)

[2] Chrysoula Terizi, Despoina Chatzakou, Evaggelia Pitoura, Panayiotis Tsaparas, Nicolas Kourtellis: Modeling aggression propagation on social media. Online Soc. Networks Media 24: 100137 (2021)

[3] A. K. Henneberger, D. L. Coffman, S. D. Gest, The effect of having aggres- sive friends on aggressive behavior in childhood: Using propensity scores to strengthen causal inference, Social development 26 (2) (2017) 295–309. doi: 10.1111/sode.12186.

URL https://www.ncbi.nlm.nih.gov/pubmed/28553013

[4] J.Lee,Y.Lee,Aholisticmodelofcomputerabusewithinorganizations,Informa-
tion management & computer security 10 (2) (2002) 57–63.

[5] A.Squicciarini,S.Rajtmajer,Y.Liu,C.Griffin,Identificationandcharacterization of cyberbullying dynamics in an online social network, in: IEEE/ACM ASONAM,
2015.