Toxicity and Networks of COVID-19 Discourse Communities: A Tale of Two Social Media Platforms

Karen DiCicco, Nahiyan B. Noor, Niloofar Yousefi, Maryam Maleki, Billy Spann, and Nitin Agarwal

University of Arkansas at Little Rock, Little Rock, AR, USA

Abstract

The issue of hateful or toxic content on social media platforms such as Twitter and Parler is on the rise and demands attention. The aim of this research is to compare and analyze toxicity between Twitter and Parler for COVID-19 discourse. Highly toxic individuals and their networks are analyzed for the two platforms. Data from January 1, 2020, to December 31, 2020, is analyzed to ascertain and compare overall network health and the evolution of toxicity over time. We found evidence that Twitter contained a higher level of toxicity regarding COVID-19 discourse than Parler. When analyzing COVID-19 vaccine discussions within the Twitter network, prominent conspiracy theory themes emerged among highly toxic users. Within the Parler COVID-19 vaccine discussion, we identified clusters of highly toxic users and important bridges aiding the spread of misinformation. These toxic conversations could impact the public health response to various non-pharmaceutical interventions (NPI's). The research demonstrates a computational method to evaluate toxicity and means for policymakers to improve the overall health of our online discourse by stemming the flow of toxicity in communities through online social networks.

Keywords

Toxicity analysis, social network analysis, COVID-19, Parler, Twitter, hate speech

1. INTRODUCTION

While major social media platforms like Facebook, Twitter, and YouTube have implemented guidelines and enforcement measures to manage toxic content and misinformation, "free speech" platforms such as Parler have been more lenient in permitting hate speech and conspiracy theories. And potentially harmful misinformation. One study showed % of active users in Parler post hateful content [1]. Parler is a micro-blogging platform that is comparable to Twitter that, by design, lacks the content moderation rules and capabilities of the platform it emulates. Parler was created prior to the emergence of COVID-19, but it has since become an important vector for online misinformation, a place where users are able to spread COVID-19 misinformation without restrictions.

The spread of abusive language and toxic content on social media can have negative impacts on communities. Analyzing toxic content provides additional insights and helps address the challenge of managing safety on social platforms. Our analysis contributes to the current research on the health of social media. For this paper, we consider misinformation to be a claim that contradicts or distorts the common understanding of verifiable facts [2].

ROMCIR 2023: The 3rd Workshop on Reducing Online Misinformation through Credible Information Retrieval, *held as part of ECIR 2023: the 45th European Conference on Information Retrieval*, April 2-6, 2023, Dublin, Ireland.

EMAIL: klcottrill@ualr.edu (A. 1); nbnoor@ualr.edu (A. 2); nyousefi@ualr.edu. (A. 3); mmaleki@ualr.edu. (A. 4); bxspann@ualr.edu. (A. 5); nxagarwal@ualr.edu. (A. 6)

In 2020, Parler, which was previously relatively unknown, experienced a sudden rise in prominence due to the efforts of conservative media personalities and politicians to distance themselves from bigger and more established social media platforms. This move was prompted by a perceived bias and censorship against conservative perspectives on these platforms. As the COVID-19 pandemic disseminated globally in 2020, both Twitter users and the predominantly far-right user community of Parler participated in conversations and shared material concerning the endeavors to vaccinate against the disease. This study undertakes a comparative assessment of the toxicity levels of COVID-19-related contents on Twitter and Parler during the duration of January 1, 2020, to December 31, 2020. We analyzed user posts for each platform and compared the evolution of toxicity levels over time. We present evidence that Twitter contained a higher level of toxicity for COVID-19 discourse than Parler over three of the four COVID-19-related content datasets we analyzed. Using the segments of the corpus that contained toxicity, we created co-hashtag graph networks for both platforms to analyze the context of the additional hashtags users were disseminating. This provided additional insight into the COVID-19 vaccine discussion within the Twitter network, which included prominent conspiracy theory themes. These included Bill Gates and the QAnon far-right conspiracy group. The graph network for Parler contained defined user communities, a misinformation echo chamber, and important bridge nodes that served to spread information throughout the rest of the network. This included a bridge node that connected an identified QAnon group to a pro-Trump group.

This work answers two research questions: 1) Does Twitter and Parler differ in terms of Toxicity on each platform? (Section 5.1) 2) Was toxic content on the COVID-19 vaccine narrative used to spread toxicity to other narratives? (Section 5.2) The remainder of this paper is organized as follows. In section 2, the related work that has been published regarding toxicity on social media is presented. Section 3 describes the data collection process and the methodology used in this paper. Section 4 presents the highlights from our results and analysis. Finally, Section 5 concludes with the contributions of this work and presents our plans and ideas for future work.

The key findings and contributions of this research are:

- Twitter contained statistically significant (p-value < 0.05) higher level of toxicity compared to Parler regarding COVID-19 discourse. (Section 5.1)
- Prominent conspiracy theory themes were identified within the Twitter network originating from the COVID-19 vaccine narrative, such as those regarding Bill Gates and the QAnon group. (Section 5.2)
- Well defined user communities with highly toxic content were identified, including a misinformation echo chamber within the Parler network. (Section 5.2)
- Significant bridge nodes were identified that spread toxic COVID-19 vaccine misinformation throughout the Parler network. (Section 5.2)

In the next section we present research from previous studies, background about existing approaches to detecting toxicity, and previous research discussing the impact of toxicity on public health discourse.

2. RELATED WORKS

Several researchers have attempted to characterize the behavior of toxic users, while also attempting to predict their future behavior. Cheng et al. investigated the long-term patterns of users displaying anti-social behavior in online forums and whether anti-social users can be identified early in their posting history using text quality metrics [3]. Guberman et al. developed a scale for assessing online aggression and applied it to a random sample of Twitter data [4]. Garimella et al. attempted to develop a technique for quantifying online discussions that cause controversy. The authors emphasized the importance of identifying these topics to understand the formation of echo chambers. They found that trolling behavior for a user decreases with the amount of time between posts, suggesting that negative behavior could have been minimized by instituting a calming period where users are unable to post comments [5]. Amrollahi (2021) discusses how users' increasing reliance on social media as a source of information can lead to filter bubbles, which can lead to polarization in society [6].

Pascual-Ferrá et al. claim that social media has an important effect on strengthening public health issues. They focused on online conversations regarding COVID-19 and wearing masks to understand toxicity's role in this discourse [7]. Majó-Vázquez et al. investigated the number of toxic conversations and patterns they follow on social media during the COVID-19 pandemic and the health of online discussions in social media [8]. In a similar study, Xue et al., by analyzing tweets shared on Twitter regarding COVID-19, investigated the discourses, sentiments, and concerns on social media [9].

Researchers have developed multiple methods and models to detect toxicity in online text. Watanabe et al. proposed a machine-learning method to detect hate speech on Twitter using sentiment and semantic-based features [10]. Gunasekara and Nejadgholi trained a multi-label classifier to detect toxicity in online conversational text, concluding that character-level text representation techniques were superior in performance than word-level representations [11]. A few studies have assessed the performance and generalizability of available toxicity detection models. Hanu, L. developed the Detoxify model which is a trained model designed to predict toxic contents. This model is capable of detecting various types of toxicity such as threats, obscenity, insults, and identity hate. The output indicates different scores for each category, based on the score the content will be labeled as toxic or not [12]. By using this method, Noor et al. detected toxicity score and different types of it. They compared the level of toxicity in three different social media platforms (Twitter, Parler and Reddit) in discussions related to COVID-19 [13].

In a study by Obadimu et al. a Non-negative Matrix Factorization (NMF) technique as an optimization problem was used to forecast commenter toxicity on YouTube. Their findings showed that the NMF model performed more accurately than other models in forecasting toxicity scores and had better computation time [14]. In another study, Obadimu et al. developed an epidemiological model to evaluate the spread of toxicity on YouTube. They used an STRS (Susceptible, Toxic, Recovered, Susceptible) model to show the similarity between the propagation of toxicity on YouTube and the spread of a disease in a population [15]. Several researchers have analyzed online toxicity from a case study perspective. Qayyum et al. analyzed the patterns of political discourse in Pakistan and India, finding that toxicity is prevalent from all sources studied [16]. In another study, Obadimu et al. evaluated five different forms of toxicity between the comments posted on pro- and anti-NATO channels on YouTube. Their analysis demonstrated that comments on pro-NATO channels are less toxic than those on anti-NATO channels [17]. Obadimu et al. analyzed toxic ideas related to COVID-19 and users who spread them on YouTube. They used social network analysis to find the influential and the top users in the network. They also applied toxicity analysis to evaluate the health of the network [18]. Pascual-Ferrá et al. evaluated the role of toxicity on Twitter regarding wearing face masks during the COVID-19 pandemic. Their results showed that tweets that used pro-mask hashtags were significantly less likely to use toxic comments while those with anti-mask hashtags were somewhat more toxic [19]. Chandrasekara et al. discussed the concept of social influence on social networks, stressing that, although there are multiple constructs involved in the social influence process, an important boundary condition involves "the direct vs. indirect peer influence" wherein influence can arise both from a user's immediate neighbor nodes (direct), but also from the common neighbors of their peers (indirect or bridge nodes) [20]. Trinkle et al. discuss how actions (sanctions) taken against actors who engage in deviant behaviors affect deterrence. Although the authors' case study involves a real-world social network in the form of employees, their results can be applied to online social networks, arming platform administrators with effective knowledge to formulate strategies for neutralization [21].

3. Data Collection

The data from both Twitter and Parler analyzed in this work consists of a corpus of user posts collected based on a list of seed hashtags related to COVID-19 from January 1, 2020, through December 31, 2020 (Table 1)

Table 1 -Hashtags used for Twitter data collection and Parler dataset filtration.

Category	Hashtags/Keywords	Records
COVID	#f*ckyourcovid, f*ckthecovid, #f*ckcovid	44,492
Lockdown	#f*ckyourlockdown/s, #f*ckthelockdown/s, #f*cklockdown/s	7,437
Mask	#f*ckyourmask/s, #f*ckthemask/s, #f*ckmask/s	28,588
Vaccine	#f*ckyourvaccine/s, #f*ckthevaccine/s, #f*ckvaccine/s	6,538

Eight datasets were created, four for each platform with corresponding hashtags and keywords. For Parler data, we used an open dataset created by Aliapoulios et al. which was a complete dataset of all Parler data from August 2018 to when Parler was shut down in January 2021 [22]. The Twitter Developer API was used to collect data from Twitter for the hashtags in (Table 1) post-hoc. Because of this, tweets and accounts removed from Twitter for being labeled misinformation were not collected. Data collected in the study will be made available upon request according to data sharing guidelines of Twitter and Parler. In the next section, we discuss our methodology and approach to calculating toxicity.

4. Methodology

Prior to executing the toxicity analysis, we removed our seed keywords and hashtags from each record in the datasets so their presence would not influence the calculated toxicity score. Non-English posts for Parler and Twitter were removed as more than 90% of our data are in English and Detoxify model can generate toxicity score effectively for English text.

We computed toxicity scores for each Parler post and Twitter tweet in the dataset using Detoxify. Detoxify is a model created by Unitary AI (https://github.com/unitaryai/detoxify) that uses a Convolutional Neural Network that is trained with word vector inputs to determine whether the text could be perceived as “toxic” to a discussion. Given a text input, the Detoxify API returns a probability score between 0 and 1, with higher values indicating a greater likelihood of the toxicity label being applied to the text. Since toxicity scores are based on a probability score of 0 to 1, toxicity scores of 0.5 or greater indicate a piece of text labeled as "toxic". Detoxify returns seven categories of toxicity scores in terms of level and type: 1) toxicity, which is the overall level of toxicity for a piece of text, 2) severe toxicity, 3) obscene, 4) threat, 5) insult, 6) identity attack and 7) sexually explicit. The reason for using Detoxify is it is an open-source comment detection python library that identifies harmful and inappropriate texts online. This is a multilingual model that has been trained in English, French, Italian, Spanish, Russian, Turkish, and Portuguese. Although it can predict toxicity by providing a score, it is not effective while there are some words related to swearing, insults, or profanity present in the text. They may predict a non-toxic text as toxic if there are certain words However, this inefficiency level is very low, and we can ignore this as it is same for both platforms. For comparison, we also explored using Google’s Perspective API, which is a related type of model with similar outputs used for determining toxicity and the scores were similar.

For our co-hashtag social network analysis, we used NetworkX (https://networkx.org/), a Python library for creating and analyzing network graphs, to generate co-hashtag graph networks for the Twitter and Parler vaccine category datasets. We used the Girvan-Newman algorithm to identify distinct communities within each network [23]. We removed low toxic posts less than 0.5 to focus on the analysis of highly toxic content [18]. The next section discusses our analysis and results of these Twitter and Parler datasets.

5. Analysis and Results

In this section, we discuss our analysis and results. First, we discuss the overall posting frequency of our seed hashtags (and keywords) and the results of our toxicity analysis. This is followed by a discussion of our social network analysis using co-hashtag graph networks and present visualizations of some of the most interesting highlights from our findings within each discussion.

5.1 Toxicity Analysis

The analysis was conducted using Twitter data, and the hashtags used as seeds were initially seen in March 2020. From March to December 2020, the COVID category had the highest number of posts compared to other Twitter datasets. The number of tweets showed a peak in mid-April, near the end of June, and a significant increase in mid-November (Figure 1).

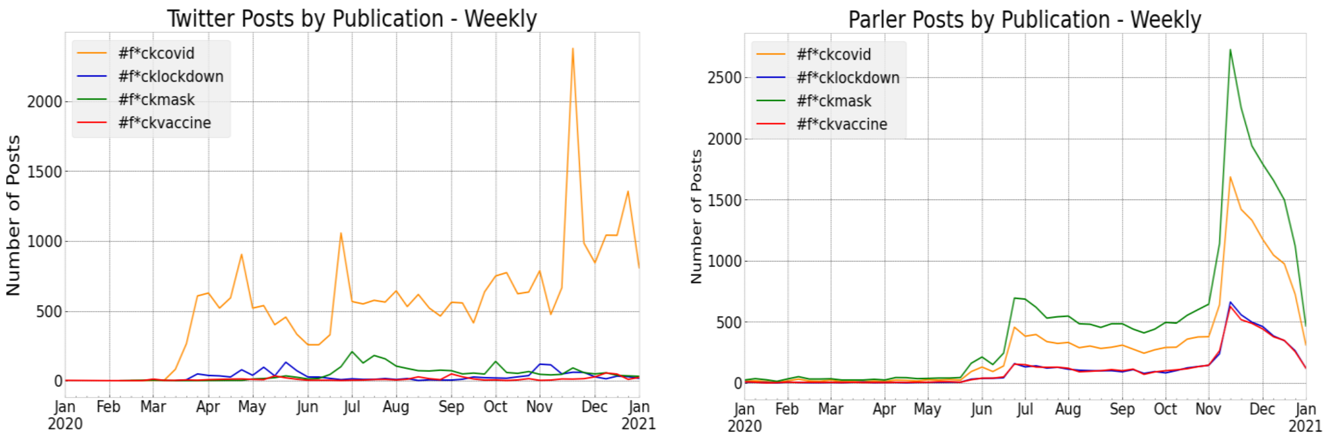

Figure 1. Twitter (left) and Parler (right) posts by weekly publication showing posting frequency.

It wasn't until the end of May that posting activity using the seed keywords was observed on Parler, as shown in Figure 1 (right). In November, there was a significant increase in posting frequency across all Parler datasets, which initially appeared to be indicative of artificial behavior. However, upon further inspection of the dataset, it was discovered that Parler users tended to use all four seed hashtags within a single post, unlike Twitter users. Our analysis revealed that there were differences in the presence of toxicity (toxicity score > 0.5) in user-generated text content between Twitter and Parler from January 1, 2020, to December 31, 2020.

When comparing the amounts of harmful content on each site, Twitter had a greater proportion of toxic tweets overall (Figure 2). This means that the majority of the Twitter content had a higher probability of being labeled as toxic than compared to the Parler content. Surprisingly, in the overall toxicity category, the Twitter content for all datasets had a higher percentage of content with toxicity scores greater than 0.7 and greater than 0.9 than did the Parler content (see Table 2). Again, Parler only exceeded Twitter with regard to the percentage of toxic content for the COVID dataset. This is an interesting result because we expected to see more toxic content on Parler due to the free-speech nature of the platform and how they tout claim their lack of censorship as a selling point for users. We also looked at the “obscene” and “insult” toxicity categories for each tweet and post for all datasets. Of the seven categories of toxicity scores obtained from Detoxify, only three contained enough data to warrant inclusion in the discussion: toxicity (overall), obscene, and insult. More Twitter content fell into the obscene category than Parler content for all datasets, with the highest percentage being within the Lockdown dataset (28.6% versus 13.06%). However, more Parler content fell into the insult category than Twitter content for the COVID dataset (18.01% versus 10.93%). The percentage of toxic content (overall toxicity category) within the Vaccine datasets varied considerably between platforms (30.98% for Twitter versus 11.93% for Parler).

Table 2. Number and percentage of toxic posts on Twitter and Parler for all eight

		Total	Percentage of Post with Toxicity Score > 0.5			Percentage of Posts with Toxicity score > 0.7			Percentage of Posts with Toxicity Score > 0.9		
Dataset	**Platform**	**Tweets/ Posts**	**Toxicity**	**Obscene**	**Insult**	**Toxicity**	**Obscene**	**Insult**	**Toxicity**	**Obscene**	**Insult**
#f*ckcovid	Twitter	28131	24.08%	22.05%	10.93%	21.65%	19.89%	8.60%	17.00%	10.06%	6.93%
#f*cklockdown		1472	34.31%	28.60%	16.37%	27.45%	22.96%	11.35%	20.11%	14.54%	6.05%
#f*ckmask		2423	31.24%	23.15%	19.81%	27.90%	20.59%	16.05%	22.86%	15.44%	5.94%
#f*ckvaccine		610	30.98%	23.28%	19.51%	27.21%	19.84%	14.75%	20.98%	14.26%	7.21%
#f*ckcovid	Parler	16361	30.51%	20.61%	18.01%	29.08%	19.16%	15.82%	15.48%	10.43%	7.03%
#f*cklockdown		5956	18.11%	13.06%	12.14%	17.45%	12.98%	11.90%	13.73%	8.14%	8.04%
#f*ckmask		26165	26.80%	15.71%	17.12%	23.38%	13.28%	13.48%	14.64%	9.96%	7.18%
#f*ckvaccine		5928	11.93%	9.06%	5.36%	10.90%	8.92%	4.82%	10.37%	8.52%	4.28%

Overall, the toxicity analysis revealed that Twitter was more toxic than Parler in all but one case, the COVID dataset. The toxic content was more obscene and insulting for both platforms. But the toxic content on Twitter was obscener than that of Parler, especially within the Lockdown dataset. The toxic content on Parler was more of an insulting type within the COVID dataset.

Next, we wanted to test the statistical validity of our findings. We conducted statistical significance testing between Twitter and Parler, using a t-test for the 4 datasets. The null hypothesis of the t-test is that the means of two groups are the same. P values for these t-tests are shown in Table 3. The p values for all these tests are significantly lower than 0.05 which implies the null hypothesis can be rejected for all four pairs. So, the alternative hypothesis is accepted which is there are significant differences between the mean toxicity scores for Twitter and Parler for all four datasets.

Table 2 reveals that there are notable differences in the mean and median toxicity values for the different platforms and contexts on the two platforms, which is a peculiar characteristic of this data. This reveals that the distributions of toxicity for these datasets are not uniform and are instead highly skewed. This shows that there are a lot of observations with very low toxicity and a few with extremely high toxicity, which is driving up the mean but is not having an influence on the median.

Table 3- Statistical Analysis of eight datasets

Pairwise Comparison	p-value <0.05	Platform	Records	Mean	SD
Twitter COVID dataset - Parler COVID dataset	2.81e-55	Twitter	28,131	0.234	0.388
		Parler	16,361	0.294	0.402
Twitter Lockdown dataset - Parler Lockdown dataset	1.87e-43	Twitter	1,472	0.326	0.406
		Parler	5,965	0.176	0.361
Twitter Mask dataset - Parler Mask dataset	5.70e-09	Twitter	2,423	0.313	0.416
		Parler	26,165	0.264	0.388
Twitter Vaccine dataset - Parler Vaccine dataset	5.05e-42	Twitter	610	0.302	0.411
		Parler	5,928	0.119	0.304

The Twitter data, for example, shows that there are a few conversations that are very toxic, and those few highly toxic conversations are driving up the overall toxicity level of the platform. This has important implications for platform administrators, who may be able to significantly reduce the strongest drivers of toxicity by moderating the relatively few, highly toxic users, rather than attempting larger platform-wide changes to all users. Table 3 also gives a good summary comparison and contrast between the two platforms in terms of the context of discussions, especially in those of vaccine and lockdown, for which Twitter is clearly more toxic. The toxicity standard deviation metrics revealed some additional unique contrasts between the two platforms (Table 3). The standard deviation of toxicity values for content within the lockdown, mask, and vaccine categories are higher on Twitter than on Parler, indicating that there is more variation in toxicity for these datasets, although values were higher for Parler for content within the COVID category. Since the percentage of toxic content in the vaccine datasets contrasted considerably between Twitter and Parler, we next drill down into those datasets and look at their network structures to identify some possible explanations for this drastic difference.

The 5-point statistical analysis in Figures 3 to 6, shows that in general, the term “f*ckcovid” on Parler is more toxic than on Twitter. However, for the other negative terms related to COVID, Twitter is more toxic. We use the top 3 most severe classes to compare in our statistical analysis. For ‘f*cklockdown’ and ‘f*ckvaccine’ hashtags, Parler is less toxic than Twitter if we consider the mean toxicity of from the boxplot for both platforms. On the other hand, for ‘f*ckcovid’ and ‘f*ckmask’ hashtags there is a significant increase in toxicity in Parler. On Twitter, the most toxic hashtag category is ‘f*ckmask’ whereas for Parler it is ‘f*ckcovid’.

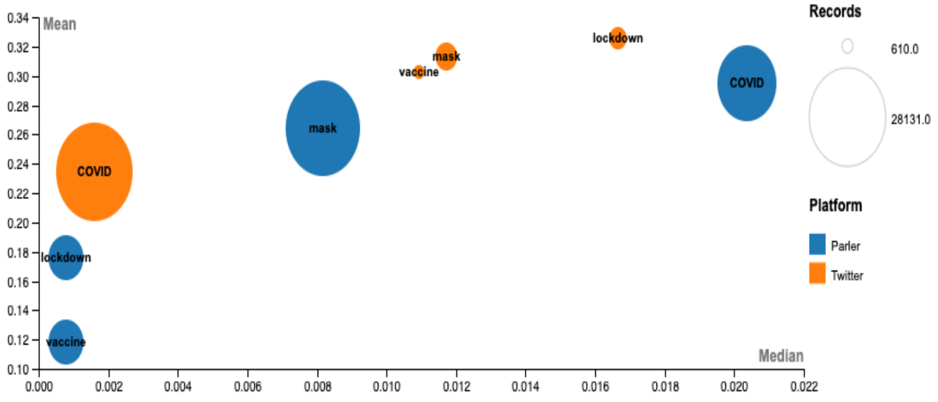

Figure 2. Platform Comparison of Toxicity Means, Medians and Standard Deviations

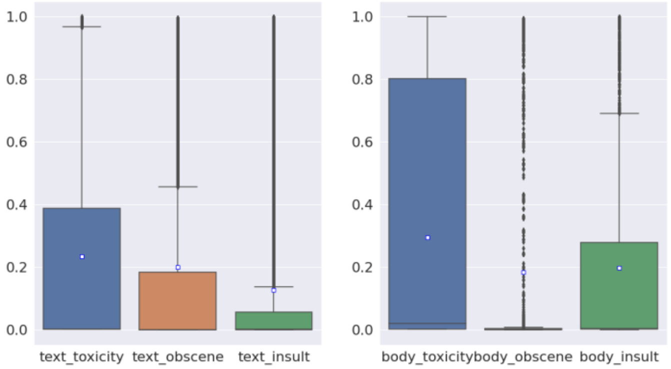

Figure 3. f*ckcovid Hashtag for three classes (Toxicity, Obscene, Insult) for Twitter(left) vs Parler

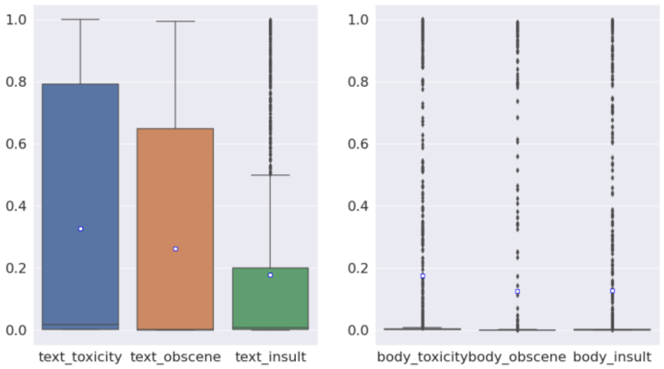

Figure 4. f*cklockdown Hashtag for three classes (Toxicity, Obscene, Insult) for Twitter(left) vs Parler

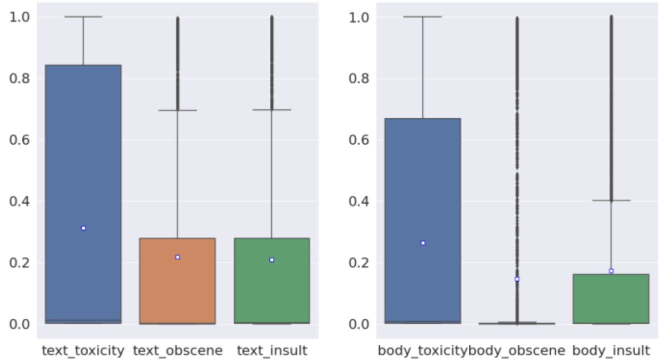

Figure 5. f*ckmask Hashtag for three types of classes (Toxicity, Obscene, Insult) for Twitter(left) vs Parler (right)

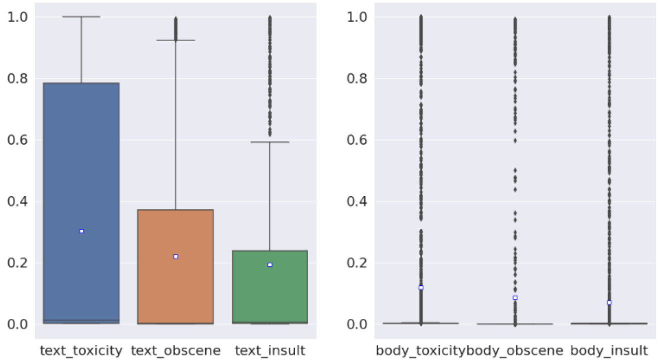

Figure 6. f*ckvaccine Hashtag for three types of toxicity (Toxicity, Obscene, Insult) for Twitter(left) vs Parler (right)

5.1 Social Network Analysis

Conducting social network analysis allows us to identify some important characteristics of the users within these datasets for each platform. Using NetworkX, we created co-hashtag network graphs to compare the Twitter and Parler Vaccine datasets filtered down to focus on the tweets and posts that scored greater than 0.5 on toxicity. This allowed us to identify the user communities, look at the context of the hashtags, and see what other topics and information toxic users shared and actively associated with the Vaccine hashtags/keywords. Our results show that the overall structure of the Twitter and the Parler co-hashtag networks vary considerably, as do the structure of their internal components. At the highest level, the Twitter co-hashtag network appears to be unstructured and somewhat scattered out with a few small clusters of users (Figure 7, left); whereas the Parler co-hashtag network is clustered and shows clear connective bridges and communities of users (Figure 7, right).

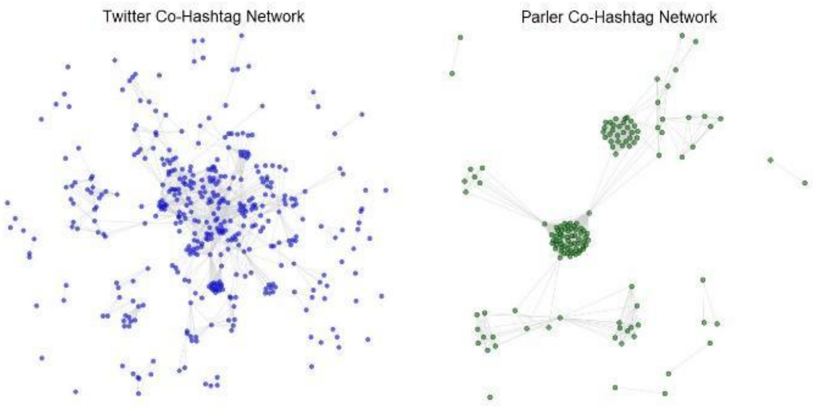

Figure 7. Toxic Twitter Co-Hashtag Network Graph (left) and Parler Co-Hashtag Network Graph (right)

Upon drilling down into these co-hashtag networks, we can identify some of the contexts within these clusters of users. When analyzing the Twitter network, we identified a mass of users sharing hashtags indicative of various Bill Gates conspiracies, anti-vaccination ideas, and the far-right conspiracy group QAnon (Figure 8). QAnon is a movement that has followers that spread false information on a variety of topics, including COVID-19 [24]. On Parler, QAnon followers often use the #wwg1wga hashtag in posts. The hashtag #wwg1wga is an abbreviation for the phrase, "where we go one, we go all" [25]. This hashtag was identified as being a clear bridge node in the Parler network, meaning that it is an important connector node and conduit by which other information flows within and throughout the network (Figure 9).

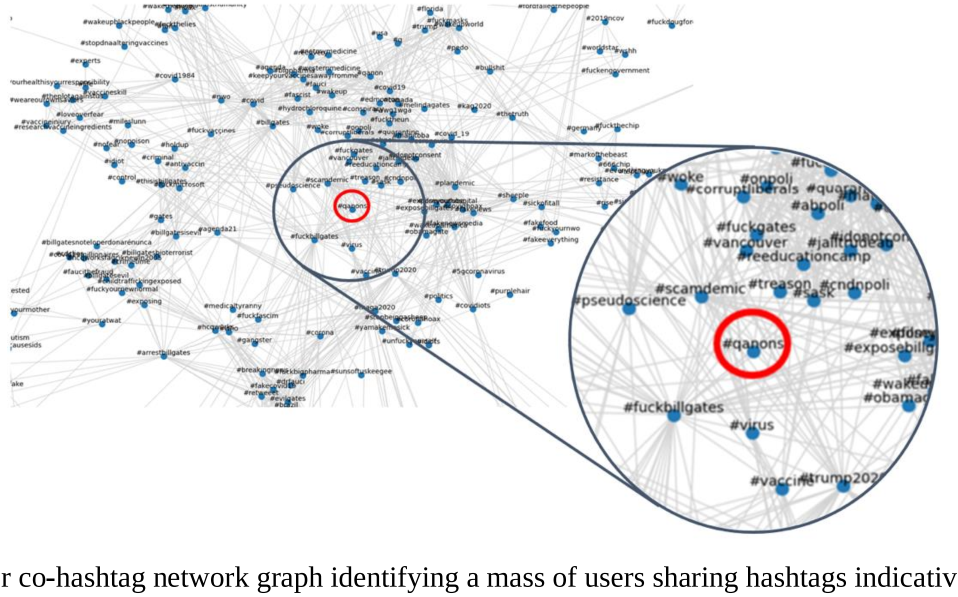

Figure 8. Twitter co-hashtag network graph identifying a mass of users sharing hashtags indicative of various Bill Gates conspiracies, anti-vaccination ideas, and the far-right conspiracy group QAnon.

The #wwg1wga bridge node can be seen connecting election conspiracy theorists (Figure 10, right), who was actively associating with the #f*ckvaccine hashtag with other hashtags indicative of the 2020 U.S. presidential election such as #maga, #trump2020, and #stopthesteal, which is related to the misinformation narrative regarding wide-spread election fraud. The other bridge nodes identified in the Parler network were #sheep and #fightback. Our drill-down also allowed us to identify a misinformation echo chamber operating within this Parler co-hashtag network (Figure 10).

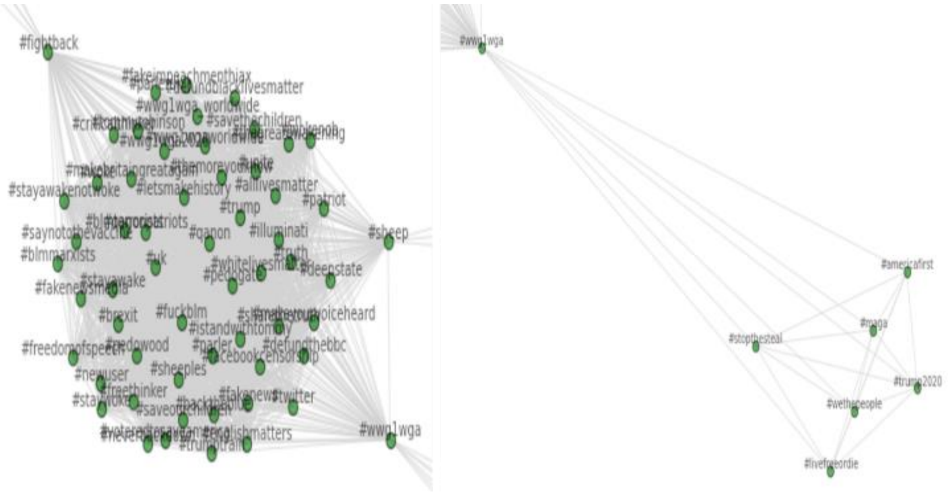

Figure 9. Parler co-hashtag network graph component identifying important bridge nodes.

This echo chamber is completely separated from the rest of the network components. Instead of the users within it exchanging new information with one another, they are repeatedly sharing the same ideas, such as #plandemic, #masksmakeyousick, #vaccineskill, and #coronavirushoax. The #plandemic hashtag was used after the "Plandemic" movie was posted online, which spread numerous conspiracy theories about COVID-19 [26]. Overall, the Twitter network showed prominent conspiracy themes such as those regarding Bill Gates and QAnon. In contrast, the Parler network showed defined user community clusters, the most concerning consisting of a misinformation echo chamber. Also, within the COVID-19 vaccine discussion within the Parler network, important bridge nodes were identified that serve to spread information throughout the rest of the network.

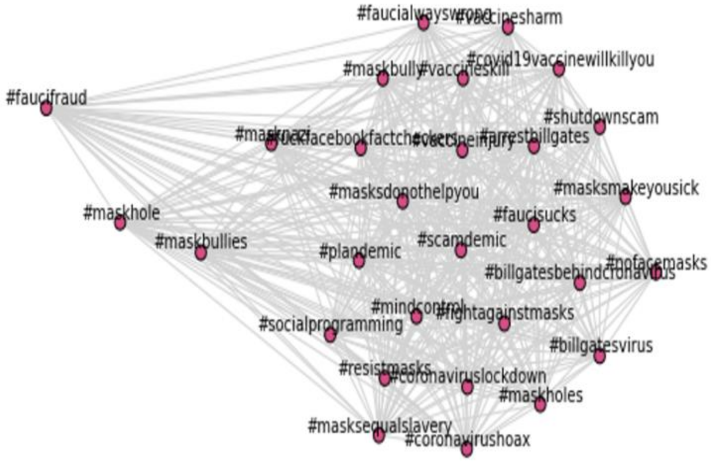

Figure 10. Parler co-hashtag network graph component identifying a misinformation echo chamber.

One interesting observation was the bridge node that connected the QAnon group connected with an identified pro-Trump group. In the next section, we discuss our conclusions and ideas for future work.

6. Conclusions and Future Work

In the specified timeframe, both Twitter and Parler exhibited considerable degrees of toxicity concerning COVID-19 topics. This study examines and contrasts the degree of toxicity and network patterns of these toxic communities on both platforms. The methods were applied to different datasets for Twitter and Parler. The results obtained from this research are preliminary, and they are restricted to a particular data arrangement. The findings indicate toxicity levels were higher overall on Twitter for all datasets except for the COVID category. It was unexpected to observe higher toxicity levels on Twitter since it is stringent content guidelines and moderation policies. Conversely, Parler's guidelines highlight a lack of moderation. One possible explanation for the unexpectedly high toxicity on the Parler COVID dataset is that Twitter began removing users and posts sharing COVID-19 misinformation in April 2020, sparking anger and prompting many users to migrate to Parler instead [27]. In addition to being detrimental to the overall health of social networks, the moderate proportion of toxic content on these platforms surrounding COVID-19 topics may affect users' perceptions of the effectiveness and importance of periodic lockdowns, wearing face masks, and becoming vaccinated. The contributions of this work include 1) Evidence that Twitter contained a higher level of toxicity regarding COVID-19 discourse than Parler; 2) When analyzing COVID-19 vaccine discussion within the Twitter network, prominent conspiracy theory themes were identified, such as those regarding Bill Gates and the QAnon group; 3) For COVID-19 vaccine discussions within the Parler network, defined clusters of users were identified, including a misinformation echo chamber; 4) In Parler, important bridge nodes were identified that spread COVID-19 vaccine misinformation throughout the rest of the network. Of specific interest was a bridge node that connected the QAnon group with an identified pro-Trump group.

The approach employed to gather and scrutinize data via the chosen seed hashtags is a possible constraint of this paper. The classification model used in this paper encounters challenges in distinguishing the intended meaning of profanity in the semantic context. Consequently, it frequently categorizes obscene language as toxic, regardless of the user's intent. In our future research, we will be mindful of this limitation and account for it accordingly. In future work, we plan to create and compare Twitter and Parler mentions, shared URL, and retweet/echo networks. These additional analyses will improve our ability to identify misinformation/conspiracy theories and to identify the users and communities that spread them. The other concerns that we can pursue for future work are users who are suspicious of being a bot. By using Botometer, we are able to detect the users who have the highest probability of being a bot. So by eliminating them we can build our network again, calculate the toxicity and expand our analysis. We will also further explore the vaccine and lockdown topics due to their exhibiting notably higher toxicity. We will also expand the analysis with the addition of topic modeling and models for the diffusion of information on OSNs. These efforts provide an important perspective on the effects of the differences between platform moderation and are the first step in a cross-platform analysis of toxicity with implications for public health and public trust.

7. Acknowledgments

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.

8. REFERENCES

[1] Israeli, A., & Tsur, O. (2022, July). Free speech or Free Hate Speech? Analyzing the Proliferation of Hate Speech in Parler. In Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH) (pp. 109-121).

[2] Guess, A., and Lyons, B., 2020. “Misinformation, Disinformation, and Online Propaganda,” Social Media and Democracy September, pp. 10–33, https://doi.org/10.1017/9781108890960.003.

[3] Cheng, J., Danescu-Niculescu-Mizil, C., and Leskovec, J., 2015. “Antisocial behavior in online discussion communities,” in Proceedings of the 9th International Conference on Web and Social Media, pp. 61–70.

[4] Guberman, J., Schmitz, C., and Hemphill, L., 2016. “First steps in quantifying toxicity and verbal violence on Twitter,” in Proceedings of the ACM Conference on Computer Supported Cooperative Work, pp. 277–280, https://doi.org/10.1145/2818052.2869107.

[5] Garimella, K., De Francisci Morales, G., Gionis, A., and Mathioudakis, M., 2017. “Reducing Controversy by Connecting Opposing Views,” in Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, Association for Computing Machinery, New York, NY, pp. 81–90., https://doi.org/10.1145/3018661.3018703.

[6] Amrollahi, A., 2021. “A conceptual tool to eliminate filter bubbles in social networks,” Australasian Journal of Information Systems (25).

[7] Pascual-Ferrá, P., Alperstein, N., Barnett, D. J., & Rimal, R. N. (2021). Toxicity and verbal aggression on social media: Polarized discourse on wearing face masks during the COVID-19 pandemic. Big Data & Society, 8(1), 20539517211023533.

[8] Majó-Vázquez, S., Nielsen, R., Verdú, J., Rao, N., de Domenico, N., & Papaspiliopoulos, O. (2020). Volume and patterns of toxicity in social media conversations during the COVID-19 pandemic.

[9] Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y., & Zhu, T. (2020). Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach. Journal of medical Internet research, 22(11), e20550.

[10] Watanabe, H., Bouazizi, M., and Ohtsuki. T., 2018. “Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection,” IEEE Access 6, pp. 13825–13835, https: //doi.org/10.1109/ACCESS.2018.2806394.

[11] Gunasekara, I., and Nejadgholi, I., 2019. “A Review of Standard Text Classification Practices for Multi-label Toxicity Identification of Online Content,” pp. 21–25, https://doi.org/10.18653/v1/w18-5103.

[12] Hanu, L. (2020). Unitary team. Detoxify. Github.

[13] Noor, N. B., Yousefi, N., Spann, B., & Agarwal, N. (2023). Comparing Toxicity Across Social Media Platforms for COVID-19 Discourse. The Ninth International Conference on Human and Social Analytics HUSO 2023 - arXiv preprint arXiv:2302.14270.

[14] Obadimu, A. Mead, E., Hussain, M., and Agarwal, N., 2019. “Identifying toxicity within YouTube video comment,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics,) pp. 214–223, https: //doi.org/10.1007/978-3-030-21741-9_22.

[15] Obadimu, A., Mead, E., Maleki, M., and Agarwal, N., 2020. “Developing an Epidemiological Model to Study Spread of Toxicity on YouTube,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 266–276, https://doi.org/10.1007/978-3-030-61255-9_26.

[16] Qayyum, A., Gilani, Z., Latif, S., and Qadir, J., 2019. “Exploring Media Bias and Toxicity in South Asian Political Discourse,” in Proceedings of International Conference on Open Source Systems and Technologies, pp. 1–8, https://doi.org/10.1109/ICOSST.2018.8632183.

[17] Obadimu, A., Mead, E., and Agarwal, N., 2019. “Identifying latent toxic features on YouTube using non-negative matrix factorization,” SOTICS 2019: The Ninth International Conference on Social Media Technologies, Communication, and Informatics.

[18] Obadimu, A., Khaund, T., Mead, E., Marcoux, T., and Agarwal, N., 2021. “Developing a socio-computational approach to examine toxicity propagation and regulation in COVID-19 discourse on YouTube,” Information Processing and Management, pp. 58, https://doi.org/10.1016/j.ipm. 2021.102660.
[19] Pascual-Ferrá, P., Alperstein, N., Barnett, D., and Rimal, R., 2021. “Toxicity and verbal aggression on social media: Polarized discourse on wearing face masks during the COVID-19 pandemic,” Big Data and Society, pp.8, https://doi.org/10.1177/20539517211023533.
[20] Chandrasekara, D., Sedera, D., & Gao, C., 2021. “Determining Boundary Conditions of Social Influence for Social Networks Research,” Australasian Journal of Information Systems (25).
[21] Trinkle, B. S., Warkentin, M., Malimage, K., and Raddatz, N., 2021. “High-risk deviant decisions: does neutralization still play a role?.” Journal of the Association for Information Systems (22:3).
[22] Aliapoulios, M., Bevensee, E., Blackburn, J., Bradlyn, B., De Cristofaro, E., Stringhini, G., and Zannettou. S. 2021. “A Large Open Dataset from the Parler Social Network,” Zendo, https://doi.org/10.5281/zenodo.4442460.
[23] https://networkx.guide/algorithms/community-detection/girvan-newman/
[24] Roose, K., 2021. “What Is QAnon, the Viral Pro-Trump Conspiracy Theory?,” NY Times, https: //www.nytimes.com/article/what-is-qanon.html.
[25] Rahn, W., and Patterson, D., 2021. “What is QAnon? What does WWG1WGA mean? The conspiracy theory that explains everything and nothing,” CBS News, https://www.cbsnews.com/news/what-is-the-qanon-conspiracy-theory/.
[26] McDonald, S., 2021. “New ‘Plandemic’ Video Peddles Misinformation, Conspiracies,” FactCheck, https://www.factcheck.org/2020/08/new-plandemic-video-peddlesmisinformation-conspiracies/.
[27] Peters, J., 2020. “Twitter will remove misleading COVID-19-related tweets that could incite people to engage in ‘harmful activity,” The Verge, https://www.theverge.com/2020/4/22/21231956/twitter-remove-covid-19-tweets-call-to-action-harm-5g.