Examining the Impact of Toxicity on Community Structure in Social Networks

Niloofar Yousefi, Nitin Agarwal, Karen Watts DiCicco, Md. Samin Morshed
COSMOS Research Center, University of Arkansas at Little Rock
Little Rock, AR, USA
Emails: nyousefi@ualr.edu, nxagarwal@ualr.edu, kwatts@uada.edu, mmorshed@ualr.edu

***Abstract*— Social media platforms, such as X continue to increase efforts to reduce harmful content, such as hate speech due to their impact on communities. The increase in harmful content was even more noticeable in 2020 with COVID-19 topics. This research systematically examines the impact of toxicity on the dynamics of communities on X, such as pro-vaccine, anti-vaccine COVID-19. Toxicity score calculated and social network analysis was performed to extract communities. These factors were co-analyzed to understand if the communities become more cohesive or more fractured over time with varying toxicity levels using Granger causality test. Our results demonstrate that in the pro-vaccine dataset, toxicity has a more substantial effect on community dynamics by fracturing communities as toxicity increases, whereas in the anti-vaccine dataset toxicity does not affect the community dynamics as much. These results have implications for how social media platforms can better moderate content and reduce toxicity within communities.**

Keywords- Toxicity; Community dynamics; Social network; Granger causality; Network analysis; Community structure.

I. INTRODUCTION

Social media platforms like Facebook and X (formerly Twitter) connect users globally but also facilitate the sharing of toxic content, which includes impolite or disrespectful language. Nockleby defines the hate speech as “any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic” [1]. Such content can harm community health and engagement [2]. Platforms have guidelines to manage toxic content due to its significant impact [3] [4]. With 48% of US adults getting news from social media [5]. Sometimes, information on social media can lead to real-life events, and vice versa [6]. Communities form and dissolve for various reasons, such as friends and family sharing content or strangers engaging with unfamiliar posts. Mixed communities blend these dynamics, creating networks of communication [7]. Users can expand communities by retweeting, mentioning, following, liking, or sharing content. However, this can evoke emotions in the users [8] . disagreements may lead to unfollowing and stopping the sharing of previously shared content [9], [10].

This research examines the impact of toxicity on community dynamics for COVID-19 vaccine content on X. A longitudinal analysis of anti- and pro-vaccine hashtags investigates differences between these communities. The two primary research questions (RQs) addressed are: **RQ1:** What is the role of toxicity in community dynamics? **RQ2:** Does toxic speech fracture a community or make it more cohesive?

Creating Sankey diagrams for both the anti- and pro-vaccine datasets, color-coded by average toxicity score, reveals toxicity dynamics at the community level. A Granger Causality test analyzed the impact of toxicity on community structure and average nodes. Results show that in the pro-vaccine dataset, increased toxicity significantly affects and fractures communities, while in the anti-vaccine dataset, toxicity has less impact. This difference is due to greater opinion diversity in the pro-vaccine data compared to the anti-vaccine data. This research explores how toxicity affects pro-vaccine and anti-vaccine communities, offering insights for improving online discourse and community management. It helps policymakers understand the behavioral differences between antagonistic and supportive communities.

In the following section 2 provides a background on existing research in toxicity and polarization, section 3 describes the details of our methodology. In Section 4, we present results and findings. Finally, section 5 summarizes our findings and discusses potential future work.

II. LITERATURE REVIEW

We review literature on hate speech followed by computational studies on community dynamics.

A. Hate Speech and Community Polarization

Toxic or hateful speech is common online and significantly impacts social network dynamics, particularly by shaping online communities and influencing information flow, especially when targeted at perceived out-groups [11]. The study [12] found that hateful posts spread faster and wider than non-hateful ones, and posts with picture attachments performed best, suggesting viral memes aid in spreading information. Authors in [13] used three measurements of graph structure to study the relationship between toxicity and the interconnectedness of X communities: connected components, modularity, and overall embeddedness.

Researchers use various techniques to measure and analyze social network structure and polarization. A study used social network analysis and natural language processing to study how political discussions on social media in Japan lead to echo chambers and user polarization [14]. Deitrick and Hu improved community detection in four X networks by integrating sentiment analysis and adding features to tweets [15]. In [16], the authors developed an index to evaluate

Copyright (c) IARIA, 2024.   ISBN: 978-1-68558-198-5

network polarization and used it to reduce polarization by promoting content on controversial subjects.

B. *Change in Response to Events*

Online communities, like offline ones, are constantly evolving and can shift in response to external factors that provoke strong emotions. In [17], the X social network was studied before and after the 2015 Charlie Hebdo attack, revealing that users became more emotional and negative. In [18], polarization in a Swiss social network during the 2011 federal elections was analyzed using time series and network measures, showing that polarization peaked before the election and returned to normal afterward. Authors in [19] analyzed communities and influential users on X in Slovenia during recent political changes and the Covid-19 pandemic, finding increased political polarization. In [20], the study of blog posts by female bloggers on women's rights focused on broker and bridge nodes and their impact on information flow.

III. Methodology

This section discusses the research methodology used to study coordination and dynamic of social media, through various types of networks.

A. *Data Collection*

In this study, we focused on two different datasets to perform community dynamics analysis that were collected using the X academic API to collect tweets related to COVID-19 from January 1, 2020, to June 30, 2021. The data was collected for various sets of hashtags that included subjects related to COVID-19. The collected hashtags were classified into anti-hashtags and pro-hashtags categories regarding vaccination. Some examples of hashtags collected for the anti-vaccine dataset include #VaccineKill, #nocovidvaccine, and #NoVaccineForMe, etc. For the pro-vaccine dataset, some hashtags include #vaccinecure, #getthevaccine, etc.

B. *Toxicity Detection*

Toxicity scores for each tweet in the datasets were computed using Detoxify, a model created by Unitary [21]. This model uses a Convolutional Neural Network (CNN) trained with word vector inputs to assess whether text is perceived as "toxic." The Detoxify API returns a probability score between 0 and 1, with higher values indicating a greater likelihood of toxicity. A threshold of 0.531 was set for identifying toxic tweets, balancing precision and recall as established by [13]. Texts with toxicity scores above 0.5 are labeled as 'toxic'.

C. *Community Dynamics*

The data was processed into a daily time series for analysis using NetworkX algorithms, such as the determining modularity and clustering coefficient, number of communities, and nodes. Modularity is a proposed division of that network into communities. It evaluates the quality of community division based on the presence of numerous edges within communities and the few between them [22], and is calculated according to Equation (1).

$$Q = \sum\nolimits_{c=1}^{n} \left[\frac{L_c}{m} - \gamma \left(\frac{K_c}{2m}\right)^2\right] \qquad (1)$$

Another characteristic of a network is the clustering coefficient, which measures the extent to which nodes within a graph tend to form clusters. A high clustering coefficient suggests that nodes are closely connected within clusters, with many connections among neighboring nodes. Conversely, a low clustering coefficient indicates a more dispersed network structure. Equation (2) calculates the clustering coefficient.

$$C_i = \frac{2e_i}{n_i(n_i - 1)} \qquad (2)$$

Calculated statistics included the minimum toxicity score, maximum toxicity score, maximum toxicity minus minimum toxicity score, toxicity mean, toxicity standard deviation, toxicity quantile 1, toxicity quantile 2 and toxicity quantile 3.

D. *The Granger Causality*

The Granger Causality test was used to predict one time series from another. A Python script facilitated this test, but before it could be conducted, the data were checked for stationarity using the Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests. The Granger Causality test requires both time series to be stationary; otherwise, the data were transformed to achieve stationarity. Two tests for both the anti- and pro-vaccine data were run for the daily data for the Granger Causality test. The first test consisted of the average toxicity score of the communities and the number of communities. The second test consisted of the average toxicity score and the average nodes of the communities.

E. *Sankey Diagrams and Five Point Statistical Summary*

Sankey diagrams illustrate node flow within communities and transitions between them over time. The Jaccard Similarity Index (Equation 3) measures the data similarity, with scores ranging from 0 to 1, where higher scores indicate greater similarity [23].

$$J(A, B) = |A \cap B| / |A \cup B| \qquad (3)$$

The data from the statistical analysis was used to develop the Sankey diagrams and five-point statistical summary. Ten dates were randomly selected for the anti-vaccine and pro-vaccine datasets. The largest community was used for the analysis if the sample date had more than one community with greater than two nodes. The date selected in the sample and the following three days were used for the data to create the Sankey diagrams to look at the community dynamics.

The five-point statistical summary was calculated for the communities for the sample data for anti- and pro-vaccine datasets. This included the minimum toxicity, maximum toxicity, and toxicity quantiles 1, 2 (median), and 3. The five-point statistical summary provides insight into toxicity score distribution within communities. Toxicity quantile 2 shows us the median for the data, while toxicity quantiles 1 and 3 show the spread of the toxicity scores, and the minimum and maximum toxicity create the data range.

IV. Results and findings

Correlation analysis examined weekly modularity scores, clustering coefficients, and average toxicity scores in anti- and pro-vaccine datasets. Sankey diagrams and Granger causality tests were also applied.

A. Modularity and Toxicity

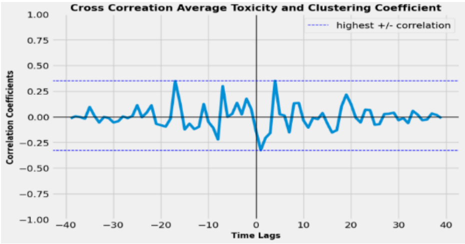
Analysis was conducted for several datasets by looking at the daily modularity scores for the user communication network (i.e., retweets and mentions) combined with the average toxicity scores, to see if toxicity is a factor in causing a community to fracture. The higher the modularity score of a network, the more modular (i.e., cohesive/well-knit) the community is. The lower the modularity score of a network, the less modular (i.e., loosely-knit) the community is. When there is a spike in toxicity in the time series, and the modularity score dips within a period of a few days, toxicity could be the cause of the fracturing of the community. A dip in modularity is noticed within a few days, because there can be a lag in the time series. The daily time series for several months of the datasets shows spikes in the toxicity mean score and dips in the modularity score the same day or day after the toxicity mean spikes. This can indicate that as toxicity rises over the network, it causes the modularity to drop, which is a sign of community fracturing.

Occurrences of significant toxicity spikes and subsequent modularity dips were observed at various points in time across the different datasets. For instance, in the anti-vaccine dataset, such occurrences were noted in January, June, and September of 2020, as well as in November 2020. In March 2021, a notable increase in toxicity was followed by a significant decrease in modularity, indicating a strong indication of community fragmentation due to increased toxicity. Similarly, even though the pro-vaccine dataset exhibited less toxicity overall—notably in January, August, October, and November 2020—the pro-vaccine results still showed frequent spikes in toxicity and corresponding dips in modularity; namely, the pro-vaccine dataset experienced mild but frequent spikes in toxicity and modularity dips, primarily occurring in April and June 2020. These findings show toxicity within a community affects modularity and has an impact on community fragmentation for both datasets.

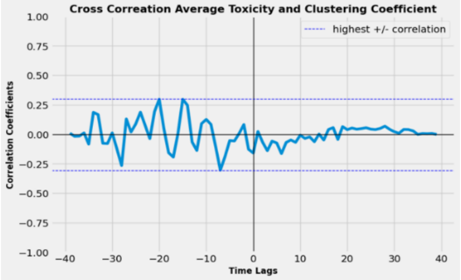
B. Clustering Coefficient and Toxicity

The analysis examined from the multiple datasets the weekly clustering coefficient of a user communication network alongside the average toxicity scores. This investigation aimed to detect whether toxicity plays a role in community fragmentation. A high clustering coefficient indicates dense connections among nodes within clusters, implying strong cohesion and frequent interactions among neighboring nodes. In contrast, a low clustering coefficient signifies a more scattered network structure, suggesting weaker ties and less frequent interactions among nodes.

Figure 1 illustrates the cross-correlation between toxicity scores and clustering coefficients for various time lags, aiming to identify the time lag at which the highest correlation between toxicity and clustering coefficient occurs. In Figure 1-A, showing the anti-vaccine dataset, a correlation is -0.23 with an 8-week lag. And Figure 1-B, representing the pro-vaccine dataset, a correlation of -0.34 is observed with a 7-week lag. This temporal aspect enriches the analysis, revealing how toxicity over time affects online community cohesion. Negative correlations (-0.23 and -0.34) indicate that, as toxicity increases, communities tend to become more fragmented. A decrease in the clustering coefficient signifies weaker member connections and reduced interaction frequency. For the anti-vaccine Figure 1-A, the correlation of -0.23 with an 8-week lag suggests that there is a modest negative association between toxicity levels and community cohesion. This means that, as toxicity increases, the community tends to become less cohesive, but this effect is observed with an 8-week delay. In contrast, in the pro-vaccine dataset Figure 1-B, the correlation of -0.34 with a 7-week lag indicates a slightly stronger negative relationship between toxicity and community cohesion compared to the anti-vaccine dataset. This implies that increases in toxicity levels are associated with more immediate and stronger decreases in community cohesion in pro-vaccine discussions, with a lag of around 7 weeks. The findings help us understand how toxicity affects the fundamental structure and behavior of online communities.



A. Cross correlation for Anti vaccine



B. Cross correlation for Pro vaccine

Figure 1. Cross correlation between toxicity and clustering coefficient.

C. Sankey Diagrams

Sankey diagrams were created to deeply investigate the dataset's community dynamics and to look specifically at the community dynamics for the anti-vaccine and pro-vaccine datasets. These diagrams help visualize what happens on the first day of a time series to the communities with an average

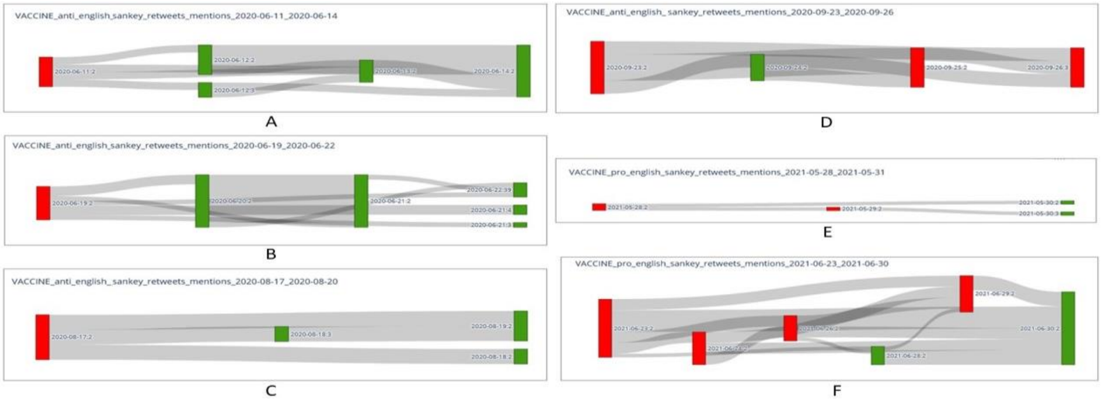


Figure 2. Community Sankey flows for different time periods.

toxicity score of greater than 0.5. Community 2020-06-11 started with two nodes that split into four different communities in the time series (Figure 2-A). The first node transitioned from the community with ID 2 on 2020-06-11 (2020-06-11:2) to the community with ID 2 on 2020-06-12 (2020-06-12:2). And the second node went to the 2020-06-12:3, 2020-06-13:2, and 2020-06-14:3 communities. These nodes' changes in community reflect the fragmentation of the original community. The tweet/retweet was the same for all nodes of the 2020-06-11:2 community, giving it the same score across the five-point statistical summary and for community 2020-06-14:2.

For the 2020-06-12:2 community, toxicity scores ranged from 0.0004 to 0.9894. In the 2020-06-12:3 community, scores ranged from 0.0007 to 0.0012. The 2020-06-14:2 community had scores ranging from 0.0004 to 0.8495. Additionally, community 2020-06-19:2 saw four out of twenty-three nodes flow into five different communities over the time series (Figure 2-B). On 2020-09-10:2, all nodes in the community retweeted the same tweet, resulting in a median toxicity score of 0.7142, with uniform quantiles and minimal variability. The 2020-06-21:2 community had a toxicity range from 0.0004 to 0.1646. On 2020-08-17:2, a community began with a median toxicity score of 0.8629, and two of its four nodes later moved to non-toxic communities (Figure 2-C). Communities 2020-08-17:2, 2020-08-18:3, and 2020-08-19:2 had identical toxicity scores due to the same retweet. The non-toxic community 2020-08-18:2 had toxicity scores ranging from 0.0004 to 0.2999. In the 2020-09-23:2 community, two nodes flowed to other communities; one node entered all subsequent communities but remained toxic, while the other moved directly to the 2020-09-26:3 community (Figure 2-D). Three additional days were analyzed, and no further connections from the 2020-09-26:3 community were found. Only the 2020-09-24:2 community had consistent five-point statistical scores. The 2020-09-23:2 community had toxicity scores ranging from 0.8765 to 0.9689, the 2020-09-25:2 community had scores between 0.7455 and 0.9842, and the 2020-09-26:3 community ranged from 0.3512 to 0.9516.

The pro-vaccine results were similar to the anti-vaccine sample dataset. In the first pro-vaccine sample dataset, the community 2021-05-28:2 started with eight nodes and then split into two different communities (see Figure 2-E). Two nodes went to other communities, while one node went straight to the non-toxic community. Two communities, 2021-05-28:2 and 2021-05-29:2, had the same score for the five-point statistical analysis. The other two communities were skewed. The 2021-05-30:2 community had a minimum toxicity score of 0.0178 and maximum toxicity score of 0.0433. The 2020-05-30:3 community had a minimum and maximum toxicity score of 0.0004 and 0.0993, respectively.

For the 2020-06-23:2 community, additional days were added to see if the toxic communities had more flowed additions. After analyzing the additional days, all the toxic communities ended in a non-toxic community for the time series (see Figure 2-F). Three of the six communities, 2021-06-24:3, 2021-06-26:2, and 2021-06-29:2, had the same score for the five-point statistical analysis. The other three communities were skewed. The 2021-06-23:2 community had a minimum toxicity score of 0.0004 and maximum toxicity score of 0.6074. The 2021-06-28:2 community had a minimum and maximum toxicity scores are 0.0004 and 0.9832, respectively. The 2021-06-30:2 community had a minimum toxicity score of 0.0004 and maximum toxicity score of 0.6074.

Overall, our analyses indicate seven out of the ten anti-vaccine communities with nodes that had connections to other communities in the time series flowed into non-toxic communities by the end of the time series for each sample. For the pro-vaccine communities, eight of the ten communities' nodes ended flowing into in non-toxic communities. This Sankey diagram analysis in Figure 2 shows that toxicity can cause the fracturing of a community.

D. Granger Causality Test

The Granger Causality test was conducted on both the anti-vaccine and pro-vaccine datasets to explore the relationship between the communities' average toxicity scores and characteristics/values, such as number of communities and number of nodes.

Initially, for the anti-vaccine dataset, the first test was between the average toxicity score of the communities and the number of communities. The ADF test was run on the toxicity and community column data, and this test was performed to assess data stationarity. For data to be considered stationary, the *p*-value must be less than 0.05. The *p*-value for the toxicity and community were 3.4056e-5 and 0.0147, respectively. So, the data series was stationary. Since the data passed the ADF test, the next step was to conduct the KPSS test to verify if the data exhibited stationarity. For the data to be stationary, the *p*-value must be greater than 0.05. The *p*-value for the toxicity and community were 0.015392 and 0.015392. So, in this case, the data series were not stationary.

Since the data did not pass the KPSS as stationary for the time series, we addressed this issue by transforming the data through differencing. After performing the differencing process, we conducted the ADF and KPSS tests again, confirming that all data series now exhibited stationarity. With both the ADF and KPSS tests passed, we proceeded to conduct the Granger Causality test using four lags. In the Granger Causality test, for either the toxicity value or the community value to Granger-cause the other variable, the p-value must be less than 0.05. The results indicated that toxicity did not Granger-cause the community values, as evidenced by p-values of 0.8127, 0.9343, 0.9898, and 0.9794 for lags 1 to 4, respectively. Similarly, the community factor did not Granger-cause toxicity for all four lags, with p-values of 0.4724, 0.7906, 0.9101, and 0.9238.

For the pro-vaccine data, the ADF test was run on the toxicity column data and the number of community column data. The toxicity data series were stationary, but the community data were not, with p-values 2.0483e-13 and 0.795, respectively. The next test performed to see if the data was stationary was the KPSS test; the toxicity, with a p-value 0.100, was stationary, but the community, with p-value 0.010, data series was not stationary. Since the community data did not pass the KPSS as stationary for the time series, the data was transformed by differencing the time series data. After the difference was performed, the ADF and KPSS tests were rerun, and all the data series passed as stationary. The Granger Causality test was performed using four lags. The outcome of the data was that toxicity does not Granger-cause the community values, and that the community values do not Granger-cause the toxicity for all four lags for toxicity and community with the p-values that are way greater than 0.05 in both cases for 4 different lags.

The second test conducted was the average toxicity score, and the average nodes of the communities.

The ADF test was run on data for the anti-vaccine dataset. The average toxicity and average nodes data series were stationary with the p-value of 3.4056e-5 and 0.049 respectively. Since the data passed the ADF test, the next test performed was KPSS to see if the data was stationary. The toxicity and average nodes data series were not stationary with p-value 0.01. Since the data did not pass the KPSS as stationary for the time series, the data was transformed by differencing the time series data. After the difference was performed, the ADF and KPSS tests were rerun, and all the data series passed as stationary. The Granger Causality test was performed using four lags. The outcome of the data shows toxicity does not Granger-cause the average nodes, as observed by p-values 0.81,0.93,0.98,0.97 for four lags.

For the pro-vaccine data, the ADF and KPSS tests were run on the toxicity column data and the average nodes column data. All the data series passed as stationary, expect KPSS for the average nodes. After the difference was performed, the ADF and KPSS tests were rerun, and all the data series passed as stationary. The Granger Causality test was performed using four lags. The outcome of the data was that toxicity does not Granger-cause for the average toxicity lag one (p-value 0.495) and four (p-value 0.07), but it does Granger-cause on lag two (p-value 0.033) and three (p-value 0.036). Toxicity does affect the average number of nodes in a community. Lag two has the strongest effect since its p-value is lower than lag three. This demonstrates that as toxicity increases, it affects the average nodes in a community with a lag, which is to be expected. One shouldn’t see Granger Cause at the same time. When looking at how the average nodes affect toxicity, the lag two, three, and four all Granger-caused. Out of all the lags, the strongest one was lag two.

V. Conclusion and Future Works

For the anti- and pro-vaccine datasets, several months show that, as the toxicity mean score rises or spikes, the modularity toxicity score decreases within a few days. When the modularity score is high and then decreases after the rise or spike of toxicity, the community becomes less tight-knit, and this shows toxic that an increase in toxicity can cause the community to fracture. Similarly, the clustering coefficient exhibits a similar trend, with an increase in toxicity corresponding to a decrease in the clustering coefficient, signifying community fracture. In the pro-vaccine dataset, an increase in toxicity leads to earlier and higher fragmentation compared to the anti-vaccine dataset. When examining community dynamics, communities starting with a toxicity score above 0.5 tend to fracture. These toxic communities often break into smaller groups, including primarily non-toxic ones. Even when members join other toxic groups, their new toxicity scores are lower than the original. Thus, a less toxic community is preferable to a highly toxic one. When the anti- and pro-vaccine sample datasets were combined, fifteen of the twenty toxic communities ended up in fully non-toxic communities by the end of the time series for those samples. This indicates that toxicity can fracture communities. The Granger Causality test on the pro-vaccine dataset revealed that toxicity affects average nodes in a community and vice versa. This may be due to greater opinion diversity in positive conversations, while negative conversations have low opinion diversity. Our results show that in the pro-vaccine dataset, increasing toxicity significantly fractures communities, whereas in the anti-vaccine dataset, toxicity has less impact on community dynamics. Other factors, such as user suspensions or disinterest in evolving topics (e.g., political discussions), can also cause communities to fracture.

This research reveals how toxicity shapes online communities, offering insights for researchers, policymakers, and community managers. By analyzing pro-vaccine and anti-vaccine discussions, it shows how toxic behavior influences community dynamics. These findings are crucial for improving online discourse and community management, helping to predict healthier communities and mitigate toxicity.

In future work, we plan to analyze a broader range of datasets from diverse sources to enhance our findings' robustness. We also aim to conduct comparative studies across multiple online platforms to further explore toxicity's impact on community dynamics.

ACKNOWLEDGMENT

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. Army Research Office (W911NF-20-1-0262, W911NF-16-1-0189, W911NF-23-1-0011, W911NF-24-1-0078), 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. 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] J. T. Nockleby, "Hate speech," *Encycl. Am. Const.*, vol. 3, no. 2, pp. 1277–1279, 2000.

[2] S. Shajari, N. Agarwal, and M. Alassad, "Commenter Behavior Characterization on YouTube Channels," in Proceedings of the eKNOW International Conference on Information, Process, and Knowledge Management, pp. 59-64, 2023.

[3] N. Yousefi, N. B. Noor, B. Spann, and N. Agarwal, "Examining Toxicity's Impact on Reddit Conversations," in *Complex Networks & Their Applications XII*, H. Cherifi, L. M. Rocha, C. Cherifi, and M. Donduran, Eds., in Studies in Computational Intelligence. Cham: Springer Nature Switzerland, 2024, pp. 401–411. doi: 10.1007/978-3-031-53503-1_33.

[4] K. DiCicco, N. B. Noor, N. Yousefi, M. Maleki, B. Spann, and N. Agarwal, "Toxicity and Networks of COVID-19 Discourse Communities: A Tale of Two Social Media Platforms," *Proc. Httpceur-Ws Org ISSN*, vol. 1613, p. 0073, pp. 30-42, 2020.

[5] M. Shaik, N. Yousefi, and N. Agarwal, "Role of Co-occurring Words on Mobilization in Brazilian Social Movements," The 30th Americas Conference on Information Systems, Salt Lake City, 2024. https://aisel.aisnet.org/amcis2024/social_comp/social_comput/20/

[6] M. Shaik, N. Yousefi, N. Agarwal, and B. Spann, "Evaluating Role of Instagram's Multimedia in Connective Action Leveraging Diffusion of Innovation and Cognitive Mobilization Theories: Brazilian and Peruvian Social Unrest Case Studies," in *2023 10th International Conference on Behavioural and Social Computing*, IEEE, pp. 1–6, 2023.

[7] S. Shajari, R. Amure, and N. Agarwal, "Analyzing Anomalous Engagement and Commenter Behavior on YouTube," The 30th Americas Conference on Information Systems, Salt Lake City, 2024.

[8] N. Yousefi, M. C. Cakmak, and N. Agarwal, "Examining Multimodal Emotion Assessment and Resonance with Audience on YouTube," In the 9th International Conference on Multimedia and Image Processing (ICMIP) ,Osaka, Japan, April 21, pp 85-93, 2024, doi: 10.1145/3665026.3665039

[9] T. C. C. Falade, N. Yousefi, and N. Agarwal, "Toxicity Prediction in Reddit," The 30th Americas Conference on Information Systems, Salt Lake City, 2024.

[10] N. Yousefi, N. B. Noor, B. Spann, and N. Agarwal, "Towards Developing a Measure to Assess Contagiousness of Toxic Tweets," in Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media, TrueHealth 2023: Workshop on Combating 13 Health Misinformation for Social Wellbeing, 5-8 June, 2023, Cyprus, p. 43.

[11] N. B. Noor, N. Yousefi, B. Spann, and N. Agarwal, "Comparing Toxicity Across Social Media Platforms for COVID-19 Discourse," In the proceeding of the Ninth International Conference on Human and Social Analytics, March 13 – 17, Barcelona, Spain, pp 21-26, 2023.

[12] B. Mathew, R. Dutt, P. Goyal, and A. Mukherjee, "Spread of Hate Speech in Online Social Media," in *Proceedings of the 10th ACM Conference on Web Science*, Boston Massachusetts USA: ACM, Jun. 2019, pp. 173–182.

[13] M. Saveski, B. Roy, and D. Roy, "The Structure of Toxic Conversations on Twitter," in *Proceedings of the Web Conference 2021*, Ljubljana Slovenia: ACM, Apr. 2021, pp. 1086–1097. doi: 10.1145/3442381.3449861.

[14] H. Takikawa and K. Nagayoshi, "Political polarization in social media: Analysis of the 'Twitter political field' in Japan," in *2017 IEEE International Conference on Big Data (Big Data)*, Dec. 2017, pp. 3143–3150. doi: 10.1109/BigData.2017.8258291.

[15] W. Deitrick and W. Hu, "Mutually enhancing community detection and sentiment analysis on twitter networks," J. Data Anal. Inform. Process. 1, 3, pp 19--29, 2013.

[16] A. Matakos, E. Terzi, and P. Tsaparas, "Measuring and moderating opinion polarization in social networks," *Data Min. Knowl. Discov.*, vol. 31, no. 5, pp. 1480–1505, Sep. 2017, doi: 10.1007/s10618-017-0527-9.

[17] S. Shaikh, L. B. Feldman, E. Barach, and Y. Marzouki, "Tweet Sentiment Analysis with Pronoun Choice Reveals Online Community Dynamics in Response to Crisis Events," in *Advances in Cross-Cultural Decision Making*, vol. 480, S. Schatz and M. Hoffman, Eds., in Advances in Intelligent Systems and Computing, vol. 480. , Cham: Springer International Publishing, 2017, pp. 345–356. doi: 10.1007/978-3-319-41636-6_28.

[18] D. Garcia, A. Abisheva, S. Schweighofer, U. Serdült, and F. Schweitzer, "Ideological and Temporal Components of Network Polarization in Online Political Participatory Media: Ideological and Temporal Components of Network," *Policy Internet*, vol. 7, no. 1, pp. 46–79, Mar. 2015.

[19] B. Evkoski, I. Mozetič, N. Ljubešić, and P. K. Novak, "Evolution of political polarization on Slovenian Twitter," *Complex Netw.*, pp. 325–327, 2020.

[20] S. T. Yuce, N. Agarwal, R. T. Wigand, M. Lim, and R. S. Robinson, "Bridging women rights networks: Analyzing interconnected online collective actions," *J. Glob. Inf. Manag. JGIM*, vol. 22, no. 4, pp. 1–20, 2014.

[21] L. Hanu and team Unitary, *Detoxify*. (Nov. 2020). Python. doi: 10.5281/zenodo.7925667.

[22] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, vol. 70, no. 6, p. 066111, Dec. 2004.

[23] Zach, "A Simple Explanation of the Jaccard Similarity Index," Statology. Accessed: Aug. 18, 2024. [Online]. Available: https://www.statology.org/jaccard-similarity/