Towards Developing a Measure to Assess Contagiousness of Toxic Tweets

Niloofar Yousefi, Nahiyan Bin Noor, Billy Spann, Nitin Agarwal

COSMOS Research Center,
University of Arkansas at Little Rock
nyousefi@ualr.edu, nbnoor@ualr.edu, bxspann@ualr.edu, nxagarwal@ualr.edu

Abstract

Toxicity is rampant on social media platforms. In this study, we use COVID-19 datasets discussing the use of face masks, vaccines, 5G, and Bill Gates conspiracy theories to examine the contagiousness of tweets with hateful or toxic content. While the majority of people believe that face masks or vaccines are effective in fighting against COVID-19, a smaller minority do not. Another major issue during the pandemic was the conspiracies surrounding 5G technology and Bill Gates. Studies indicate that the worries surrounding mobile technology have fueled certain conspiracy theories linking 5G and Bill Gates with the COVID-19 virus. Using these polarizing datasets, we measure the type and intensity of hate speech in each dataset. Then we propose a definition for a toxicity contagiousness score to study the propagation of toxicity in each dataset. Our study revealed that in the 5G dataset, low toxic content has a positive correlation with the contagiousness score. Highly toxic content in the Bill Gates and pro-mask datasets also showed a positive correlation. However, in the anti-community dataset, such as anti-mask and anti-vaccine, highly toxic content had a negative correlation with the contagiousness score. These findings shed light on how different types of content and contexts can influence the spread of toxicity in online communities.

Introduction

Social media platforms have transformed the way people interact and share information with each other. However, the use of these platforms has led to the proliferation of toxic behaviors, which poses significant challenges for both users and social media companies. The toxic content on social media focuses on "threats, obscenity, insults, and identity-based hate" and also be included harassment and socially disruptive persuasion, such as misinformation, radicalization, and gender-based violence. (1) Toxicity stifles communication and connection and makes it difficult for people to connect with each other. This problem has been amplified by the COVID-19 pandemic, which has sparked heated debates and controversies regarding face masks and the COVID vaccine. Moreover, there were a large number of conspiracy theories regarding 5G and Bill Gates spread on the internet. These theories encompass a wide range of claims, from the assertion that 5G weakens people's immune systems to the notion that 5G technology can alter people's DNA, making them more vulnerable to catching COVID-19. However, the widespread occurrence of toxic behavior related to 5G technology and Bill Gates on these platforms has become a significant problem for both users and communities. It is essential to identify toxicity and its spread in social media, as the presence of toxicity hinders users' ability to interact and freely express themselves and negatively affects the overall well-being of the online community. To address this issue, researchers have analyzed tweets to evaluate the extent of the toxicity and its propagation during the COVID-19 pandemic. This study focuses on topics such as mask, vaccine, 5G, and Bill Gates. More about keywords will be mentioned in the data collection section.

Our results indicate that in the 5G dataset, we found a positive correlation between the contagiousness score and average toxicity for low toxic content. This means that low-toxic content is more likely to be contagious and widely shared in this context. In contrast, the Bill Gates dataset shows highly toxic content is widely shared and has a strong contagiousness score. Similarly, the pro-mask dataset suggests that highly toxic content in that community is also widely shared and has a positive correlation with the contagiousness score. However, In the anti-community against masks and vaccines, highly toxic content is less contagious and has a negative correlation with contagiousness score, possibly due to differing opinions and beliefs. Next, we discuss the relevant literature.

Literature Review

Toxicity in Social Media: In several former studies, various Machine learning approaches have been used to detect toxicity in social media. Fan et al. (3) propose a model for detecting toxicity and classification in Twitter using the Bidirectional Encoder Representations from Transformers (BERT). Obadimu et al., (4) to assign a toxicity score for comments on YouTube, used Google's Perspective API. Then, by applying Latent Dirichlet Allocation (LDA), they found the topics that the toxic contents were posted. In a similar research, Noor et al., (5), aim to evaluate the level of toxicity present in discussions related to COVID-19 on Twitter, Parler, and Reddit. To determine the platform with the highest toxic content, he compares the toxicity scores across

the three social media platforms. In a similar study, DiCicco et al. (6) detected and compared the level of toxicity in two different platforms regarding COVID-19 discourse. Then they attempt to analyze the network of highly toxic users. In another study, Chakrabarty (7) proposed a model to detect various types of toxicity, such as threats, insults, obscenity, and identity-based hatred. P. A. Ozoh et al., (8) developed a model to detect toxicity by differentiating toxic contents from non-toxic contents. In this research, we implement Detoxify algorithm (12) to detect the toxicity and its various categories.

Toxicity Propagation: Obadimu et al. (9) developed a technique for toxicity detection and proposed features that propagated the toxicity contents on YouTube. Further, Obadimu et al. (4) proposed a novel approach to explain the propagation of toxicity on YouTube using epidemic methods. They concluded that toxicity propagation is comparable to the spread of a contagious disease. Mathew et al. (10) focus on the diffusion dynamics of the content of hateful users and non-hateful users and find out that the propagation of content made by hateful users is faster than by non-hateful users. In another research, Lopez-Sanchez (11) investigated hate speech propagation on social networks using agent-based modeling. They modeled three countermeasures: education, deferring hateful content, and cyber activism, and tested their effectiveness in combating hate speech. Our research investigates the propagation of toxicity by focusing on the effect of retweets, replies, and likes, which has not been reported in previous literature. Next, we describe our research methodology.

Methodology

Here, we discuss our research methodology, including data collection, toxicity assessment, and developing a toxicity propagation score.

Data Collection: Data collection is one of the critical tasks in the analysis and propagation of toxicity. Twitter Academic API and different hashtags are used to collect COVID-19 related tweets for our datasets. After analyzing some contemporary tweets' hashtags at that time, we have selected some keywords that portray the scenario of the tweets, whether it is posted by pro-community or anti-community. Our datasets for 5G, Bill Gates, Anti Vaccine, and Pro Vaccine cover the period from February 2020 to June 2021. And Anti Mask and Pro Mask extend from January 2020 to December 2020. The attributes in each dataset are Tweet (The original tweet); Created_at (Time when a tweet is posted); Retweet_count (Number of retweets); Quote_count (Number of quotes); Reply_count (Number of replies); Like_count (Number of likes). Data preprocessing steps included removing missing values, emojis, symbols, and URLs, and converting uppercase letters to lowercase. Once the preprocessing step is complete, the number of tweets available for analysis are as follows: 33,403 for the 5G dataset, 67,780 for the Bill Gates dataset, 474,534 for the anti-vaccine dataset, 601,565 for the pro-vaccine dataset, 40,674 for the pro-mask dataset, and 35,333 for the anti-mask dataset.

Toxicity Detection: This section describes the technique and method used in our research to detect toxicity. The Unbiased Detoxify Model is used to compute the toxicity score for each tweet. Detoxify is a model created by Unitary AI (https://github.com/unitaryai/detoxify) (12). Detoxify uses a Convolutional Neural Network trained with word vectors to detect the toxic language in tweets. It provides a probability score between 0 and 1 for each input. The output of the Detoxify model is a set of probabilities related to various labels that indicate the likelihood of the input text containing toxic language. These labels correspond to different types of toxicity, such as Toxic; Severe_toxic; Obscene; Threat; Insult; and Identity_hate. Therefore, the probability value indicates the likelihood that the input text contains language that falls into that category. For example, the toxic label represents the probability that the text is generally toxic. The probability score ranges from 0 to 1, with a higher score indicating a greater likelihood of the text being toxic. The tweets are classified into two groups based on a threshold for the toxicity score (4) (14).

Toxicity Contagiousness Score: There are various ways a toxic tweet can propagate. The most significant way is by retweeting the toxic post. Retweeting a post means that different users have shared the post through their account with all their followers. The more a toxic tweet is retweeted, the more the toxicity propagates. Additionally, the act of sharing a tweet again while adding one's own comment is known as quoting the tweet. Another way to accelerate the propagation of toxicity is by replying to a toxic tweet. In this case, the followers of the user who replies to the toxic tweet are exposed to toxicity. One more factor that impacts toxicity propagation is the number of likes the toxic tweet gets. Tweets with a higher number of likes are shown more than those with fewer likes (13). Developing a quantitative understanding to study the contagiousness of toxicity will give us better insight, indicating how the toxicity will spread. To this end, we propose the Toxicity Contagiousness Score (equation 1). The toxicity contagiousness score is a sum of the retweet count, quotes count, reply count, and like count of a tweet.

$$ContagiousScore = RetweetCount + QuoteCount + ReplyCount + LikeCount$$

Results

The research findings are presented in this section. First, we describe our findings from the toxicity assessment, followed by the findings from the toxicity contagiousness score.

Toxicity Assessment: As mentioned before, we used the Detoxify model to analyze tweets from our datasets, resulting in a toxicity score between 0 and 1. A threshold was set to distinguish highly toxic content from less toxic content. Saveski et al,. (14) examined a threshold level of 0.531 to differentiate between toxic and non-toxic tweets. They manually classified 3000 tweets and chose 0.531 as the threshold. In our study, we rounded the score to 0.5. In this work, posts of tweets that have a toxicity score of 0.5 or more are considered highly toxic, while anything less than 0.5 is

considered less toxic, according to Saveski et al. Our study shows that the average toxicity of highly toxic tweets is consistent across all six datasets, ranging from 0.79 to 0.88. However, due to a higher number of posts, the Pro-vaccine and Anti-vaccine datasets had a slightly lower average toxicity score of 0.79. The average toxicity levels across all tweets in the six datasets are also similar, ranging from 0.03 to 0.1. However, Number of highly toxic posts in datasets: 5G (1,799), Bill Gates (4,670), anti-vaccine (24,694), pro-vaccine (11,322), pro-mask (2,872), anti-mask (4,184). The findings show that although the number of individuals sharing highly toxic posts may be lower in these cases, the toxicity scores of their posts are still quite high. In addition, the result indicates that people share more highly toxic posts in the anti-community than in the pro-community. In general, we can mention that the anti-Mask dataset has the highest toxicity score among all the datasets. Our findings show that the average toxicity scores of all datasets are significant and comparable. The Insult subcategory had the highest scores among all categories, ranging from 0.53 to 0.64, indicating a trend among users to share highly toxic content related to insults. Additionally, the second-highest scoring category was Obscene, ranging from 0.31 to 0.43.

Contagiousness of Toxicity: In this section, we calculate the contagiousness score, which measures the spread of content on Twitter. The score is calculated by determining the number of retweets, quotes, replies, and likes for each tweet, as outlined in equation 1 from section 3. We applied MinMaxScaler from sklearn library to scale each score to the same range.

Figures 1 and 2 display the correlation between the average toxicity and the average contagiousness score, broken down by month from February 2020 to June 2021 for the 5G and Bill Gates datasets. Figure 1 focuses on less toxic content in the 5G dataset, revealing a positive correlation between the average toxicity of low toxic content and the contagiousness score. Fluctuations in the average toxicity align with corresponding fluctuations in the average contagiousness score, indicating a direct correlation between the two. As the average toxicity of less toxic content increases or decreases, the average contagiousness score does the same, demonstrating people's tendency to share non-toxic content. The correlation score between the two is approximately 0.59. Subsequently, we analyzed tweets with highly toxic content as well. The findings do not indicate a strong correlation between the contagiousness score and the average toxicity of highly toxic posts. In fact, the contagiousness score remains largely low, suggesting that people are unlikely to spread highly toxic posts that are related to 5G. This indicates that despite the presence of highly toxic content, it does not possess the contagiousness required for it to be widely shared among individuals.

In contrast to the analysis of 5G, figure 2, which is focused on highly toxic content, reveals a correlation between the average toxicity and the contagiousness score. Data from April 2020 to January 2021 shows a direct relationship between the two (correlation score close to 0.43), indicating that highly toxic content related to Bill Gates is more likely to be shared among individuals. Afterward, the less-toxic

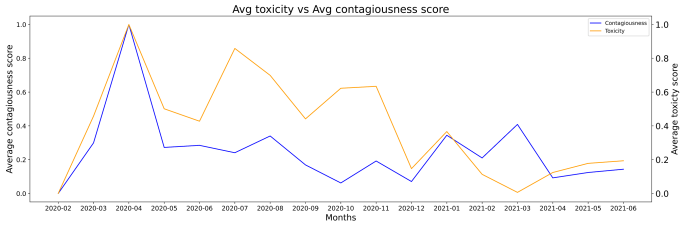

Figure 1: Average low toxicity vs. average contagiousness score for 5G

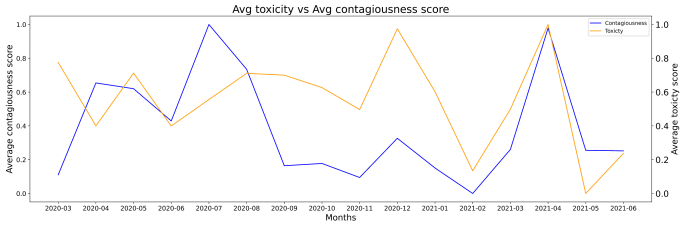

Figure 2: Average high toxicity vs. average contagiousness score for Bill Gates

contents were analyzed for the Bill Gate dataset. The results demonstrate the absence of a correlation between the contagiousness score and the average toxicity of low toxic content related to Bill Gates. With the exception of a peak in the contagiousness score observed during the two-month period from December 2020 to February 2021, the majority of the time, the contagiousness score remains low, suggesting that low toxic content related to Bill Gates is not widely shared among individuals.

Subsequently, analysis and comparison of the correlation between the contagiousness score and the toxicity score will be conducted in the contexts of both anti-community and pro-community. Figures 3 and 4 demonstrate how the contagiousness score and the toxicity score are related to one another in anti-community. Figure 3 shows that while the average toxicity score for anti-mask data decreased between June 2020 and September 2020, there was a significant increase in the average contagiousness score. However, from September 2020 to November 2020, there was a decrease in the contagiousness score as the toxicity score increased. Subsequently, figure 4 pertains to the Anti-Vaccine community. The figure illustrates that between January and February 2020, as the average toxicity score increased, the contagiousness score decreased. After that, there was not a significant change in the contagiousness score for several months until December 2020. Starting in December 2020, the COVID-19 Vaccine was released, and there were fluctuations in the toxicity score from December 2020 to June 2021. During this period, as the average toxicity score fluctuated, either increasing or decreasing, the contagiousness score changed in the opposite direction, exhibiting a -0.32 correlation.

The findings of this study suggest that in anti-communities, the contagiousness score fluctuates in the opposite direction as the average toxicity score increases or de-

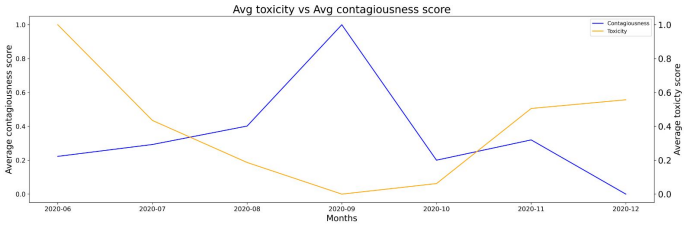

Figure 3: Average high toxicity vs. average contagiousness score for Anti-mask

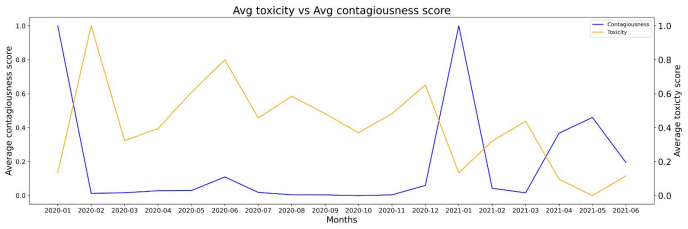

Figure 4: Average low toxicity vs. average contagiousness score for Anti-vaccine

creases. This indicates that individuals in these communities are less likely to share and spread posts with toxic content.

In figure 5, which pertains to the pro-mask dataset in the pro-community, there is a correlation between the average toxicity score and the average contagiousness score. This means that the contagiousness score changes in correspondence to the fluctuations in the average toxicity score, either increasing or decreasing accordingly. This suggests that individuals in the pro-mask community are more likely to share and spread posts that contain high levels of toxic content. To determine whether there is a link between the contagiousness of content and its toxicity level, we used a statistical test called the Granger causality test. The Granger Causality test will examine the potential dependence of the contagiousness score on the toxicity score. To ensure the validity of the Granger causality test, it is important to first determine if our data is stationary and it is not changing significantly over time. This is done by running the Augmented Dickey-Fuller (ADF) test, which helps us determine whether the data is exhibiting a stable statistical pattern over time. The test result of -19.13 indicates that there is strong evidence to support the idea that the time series data is stationary. In other words, if the ADF test statistic has a larger negative value, it provides stronger evidence that the null hypothesis of non-stationarity is incorrect. The null hypothesis is a statement we assume to be true until proven otherwise. Thus, we can proceed with applying the Granger Causality test. Following that, the basic idea of the Granger causality statistical method is that if one time series can cause another time series, then variations in the former should be able to predict changes in the latter. Our null hypothesis stated that the toxicity score does not apply a Granger-causal effect on the contagiousness score. Our findings indicate that for four distinct lags of one month, the corresponding p-values exceed 0.7, which suggests that the observed effects are unlikely to have occurred by chance. Thereby rejecting our null hypothesis. Therefore, we claim that the toxicity score within our pro-mask dataset shows a causal impact on the contagiousness score. On the other hand, in the pro-vaccine dataset, there were no significant changes observed in the contagiousness score until April 2021 to June 2021. During this period, with an increase in the average toxicity score, the contagiousness score also increased.

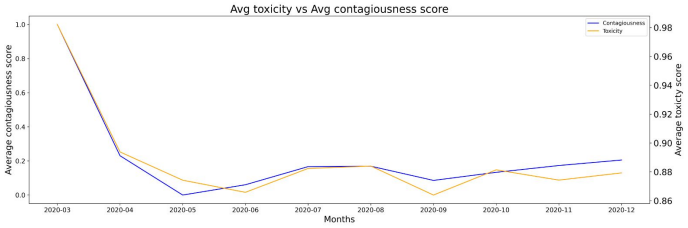

Figure 5: Average high toxicity vs. average contagiousness score for Pro-mask

Conclusion

Covid-19 led to a rise in conspiracy theories about 5G, Bill Gates, and the virus, while people disagreed on the efficacy of masks and vaccines and discussed these topics on social media. Our analysis reveals interesting insights into the correlation between the contagiousness score and average toxicity levels of different content in different datasets. The results indicate that in the 5G dataset, there is a correlation between the contagiousness score and average toxicity for low toxic content. And for the Bill Gates dataset, we observed a correlation between the average toxicity of highly toxic contents and the contagiousness score. Similarly, a positive correlation was observed in the pro-mask dataset between the contagiousness score and the average toxicity level of highly toxic content. As the toxicity score increased, the contagiousness score also increased, indicating that people were more likely to share and spread these contents. On the other hand, when the toxicity score decreased, there was a decrease in the likelihood of people sharing the content. This could be due to the fact that highly toxic content could be seen as more provocative or attention-grabbing, which could also contribute to its increased contagiousness score. Conversely, within our anti-community, which includes anti-mask and anti-vaccine, the average toxicity score and contagiousness score are negatively correlated. This means that within the anti-community, the more toxic content is on average, the less likely it is to be widely shared. In other words, toxic content that is particularly aggressive or harmful tends to be less widely shared among those who are against masks and vaccines.

Future work could involve expanding data and considering different hashtags and keywords, categorizing data into neutral, mildly toxic, and highly toxic groups, comparing the results with existing methods, applying models to other platforms like YouTube and TikTok, and using Twitter's new feature for calculating contagiousness scores.

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.

References

[1] Sheth, A., Shalin, V. L., Kursuncu, U. (2021, April). Defining and detecting toxicity on social media: Context and knowledge are key. Neurocomputing. https://doi.org/10.1016/j.neucom.2021.11.095

[2] Lang, J., Erickson, W. W., Jing-Schmidt, Z. (2021). MaskOn! MaskOff! Digital polarization of mask-wearing in the United States during COVID-19. PloS one, 16(4), e0250817.

[3] Fan, H., Du, W., Dahou, A., Ewees, A. A., Yousri, D., Elaziz, M. A., ... Al-qaness, M. A. (2021). Social media toxicity classification using deep learning: Real-world application uk brexit. Electronics, 10(11), 1332.

[4] Obadimu, A., Mead, E., Maleki, M., Agarwal, N. (2020, October). Developing an epidemiological model to study spread of toxicity on YouTube. In International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation (pp. 266-276). Springer, Cham.

[5] 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 2023, forthcoming.

[6] K. DiCicco, N. B. Noor, N. Yousefi, B. Spann, M. Maleki, and N. Agarwal.(2023) "Toxicity and networks of COVID-19 discourse communities: A tale of two media platforms." In The 3rd Workshop on Reducing Online Misinformation through Credible Information Retrieval, forthcoming.

[7] Chakrabarty, N. (2020). A machine learning approach to comment toxicity classification. In Computational intelligence in pattern recognition (pp. 183-193). Springer, Singapore.

[8] Ozoh, P. A., Olayiwola, M. O., Adigun, A. A. (2019). Identification and classification of toxic comments on social media using machine learning techniques. International Journal of Research and Innovation in Applied Science (IJRIAS).

[9] Obadimu, A., Khaund, T., Mead, E., Marcoux, T., Agarwal, N. (2021). Developing a socio-computational approach to examine toxicity propagation and regulation in COVID-19 discourse on YouTube. Information Processing Management, 58(5), 102660.

[10] Mathew, B., Dutt, R., Goyal, P., Mukherjee, A. (2019, June). Spread of hate speech in online social media. In Proceedings of the 10th ACM conference on web science (pp. 173-182).

[11] Lopez-Sanchez, M., Müller, A. (2021). On Simulating the Propagation and Countermeasures of Hate Speech in Social Networks. Applied Sciences, 11(24), 12003.

[12] GitHub. Retrieved from https://github.com/unitaryai/detoxify

[13] https://www.socialchamp.io/blog/twitter-algorithm/

[14] Martin Saveski, Brandon Roy, and Deb Roy. 2021. The Structure of Toxic Conversations on Twitter. In Proceedings of the Web Conference 2021 (WWW '21). Association for Computing Machinery, New York, NY, USA, 1086–1097. https://doi.org/10.1145/3442381.3449861