Utilizing Topic Modeling and Social Network Analysis to Identify and Regulate Toxic COVID-19 Behaviors on YouTube

Adewale Obadimu
LinkedIn Corporation
wobadimu@linkedin.com

Tuja Khaund
University of Arkansas at Little Rock
txkhaund@ualr.edu

MaryEtta Morris
University of Arkansas at Little Rock
mmoris@ualr.edu

Esther Mead
University of Arkansas at Little Rock
elmead@ualr.edu

Nitin Agarwal
University of Arkansas at Little Rock
nxagarwal@ualr.edu

Abstract

As the novel coronavirus (COVID-19) continues to ravage the world at an unprecedented rate, formal recommendations from medical experts are becoming muffled by the avalanche of toxic and abusive content posted on social media platforms. This high amount of toxic content prevents the dissemination of important and time-sensitive information and ruins the sense of community that online social networks (OSNs) seek to cultivate. In this article, we present techniques to analyze toxic content and actors on YouTube during the initial months after COVID-19 was made public. Our dataset consisted of 544 channels, 3,488 videos, 453,111 commenters, and 849,689 comments. We applied topic modeling based on latent Dirichlet allocation (LDA) to identify dominant topics and evolving trends within the comments on relevant videos. We used social network analysis (SNA) to detect influential commenters, and toxicity analysis to measure the health of the network. SNA allows us to identify the top toxic users in our network, which helped to inform experiments simulating the impact of the removal of these users. Through this work, we were able to profile toxic comments related to COVID-19 on YouTube, and the commenters who were propagating this toxicity. In addition, we devised a set of experiments in an attempt to show how if social media platforms eliminate certain toxic users, they can improve the overall health of the network by reducing the overall toxicity level.

Keywords

Toxicity Analysis, Social Network Analysis, Topic Modeling, Pandemic, COVID-19, YouTube, Social Media.

Introduction

In recent years, we have witnessed an exponential growth in the amount of digital content that is being pushed on various social media platforms. Now, more than ever, online social networks (OSNs) have become a go-to place for obtaining news, information, and entertainment. Despite the myriad advantages of utilizing OSNs, a consensus is emerging suggesting the presence of an ever-growing population of malicious actors who utilize these networks to spread toxicity and harm others. These actors (hereafter referred to as toxic users) thrive on disrupting the norms of a given platform and causing emotional trauma to other users. To set a context for our work, we give an operational definition of toxicity as "the usage of rude, disrespectful, or unreasonable language that will likely provoke or make another user leave a discussion." [1]. In this regard, toxicity analysis is different from sentiment analysis, which is the attempt to assign sentiment scores of positive, neutral, and negative to text data. OSNs like YouTube connect and stimulate interactions between people from all over the world. Unfortunately, those interactions are not always positive. YouTube provides a particularly fertile ground for toxic behavior, as most videos are

accompanied by comment sections. Unlike other OSNs, there is no 'friending' or approval process to control who may comment on a video. Anyone with a user account can freely engage in this negative and harmful behavior, which often incite further negative actions, much like the spread of a disease. While toxic behavior is an unfortunate hallmark of online interaction, certain events can increase the prevalence of this behavior. The recent pandemic due to the novel coronavirus (COVID-19) is one such event. Utilizing COVID-19 as a case study, our objectives in this paper are to identify and profile toxicity within the network, examine themes among toxic comments, investigate connections among toxic commenters, and simulate the removal of these commenters to discover the impact on the health of the network. In an attempt to meet these objectives, we outline a detailed methodology for analyzing a COVID-19 discourse dataset, and, subsequently, a technique for simulating how OSNs can improve the over-all health of their networks by eliminating highly toxic users. We present the following contributions in this paper:

1. Applied topic modeling based on LDA and Social Network Analysis to identify and visualize common themes among the toxic comments. Similar themes were found using both methods.
2. Profiled common behavior patterns among toxic users. We discovered that commenters with similar toxic levels tend to stay together and replies to toxic comments often were at a similar toxicity level to the original comment.

The remainder of this paper continues as follows. First, we highlight extant literature that are germane to our work in the section on Literature Review. We then describe our data collection and processing in the Methodology section. Next we delve into the analysis techniques applied to understand the data and our findings. We also present experimental simulations on the impact of removing toxic users in the network. Finally, we conclude with ideas for future work.

Literature Review

Prior studies have shown that online users participate in toxic behaviors out of boredom [2], to have fun [3], or to vent [4]. A comprehensive examination of various forms of toxicity was conducted by Warner et al. [6]. [11] noted that given the right condition, anyone can exhibit toxic tendencies. Another study [9] shows that toxic users become worse over time, in terms of the toxic comment they post, and they are more likely to become intolerant of the community. One of the major obstacles in understanding toxic behavior is balancing freedom of expression with curtailing harmful content [5]. Closer to our work is a study that analyzed toxicity in multiplayer online games [8]. The authors indicated that a competitive game might lead to an abuse of a communication channel. Other research [7] combined lexical and parser features to identify an offensive language in YouTube comments. Davidson et al. [12] presented a dataset with three kinds of comments: hate speech, offensive but non-hateful speech, and neither. Hosseini et al. [13] demonstrated the vulnerability of most state-of-the-art toxicity detection tools against adversarial inputs. After experimenting with a transfer learning approach, Grondahl et al. [14] concluded that hate speech detection is largely independent of model architecture. The authors showed that results are mostly comparable among models but do not exceed the baselines.

Analysis and Results

To understand the online toxicity surrounding the COVID-19 pandemic, we employed a methodology that consists of four components: 1) data crawling and processing, 2) toxicity analysis, 3) topic modeling, and 4) social network analysis. We leveraged the YouTube search Data APIv3[1] to extract channels, videos, and comments using the following keywords, coronavirus, corona, virus, COVID19, COVID, and outbreak. The time frame of our dataset spans the period from January 2020 through April 2020. To reduce noise in extracted data, several data processing steps were performed including data formatting, data standardization and data normalization. Our dataset consisted of 544 channels,3,488 videos, 453,111 commenters, and 849,689 comments (826,569 of which were unique comments). The comments were primarily from videos that were categorized as "News & Politics" (94%), followed by "Entertainment" (0.08%). We delve deeper into the dominant topics in a later section. Overall, however, in this work, we combine the techniques of topic modeling, toxicity analysis, and social network analysis to emphasize the

[1] https://developers.google.com/youtube/v3/docs/search

usefulness of how this combination of techniques can be applied to social media data to achieve the potential objective of improving the overall health of an OSN.

Toxicity Detection

The next step in our methodology was to compute toxicity scores to each comment in the dataset. To accomplish this, we leveraged a classification tool called Perspective API2, which was developed by Google's Project Jigsaw and 'CounterAbuse Technology' teams. This model uses a Convolutional Neural Network (CNN) trained with word vector inputs to determine whether a comment could be perceived as "toxic" to a discussion. The API returns a probability score between 0 and 1, with higher values indicating a greater likelihood of the toxicity label being applied to the comment.

Topic Modelling

Perceiving the discussed topic in social media is imperative to detecting various topic facets and extracting their dynamics over time. Hence, we leveraged a topic model based on latent Dirichlet allocation (LDA) [15] to automatically discover evolving topics that are related to COVID-19. In LDA models, each document is composed of multiple topics. A topic is a collection of dominant keywords that are typical representatives of an entire corpus. In our case, each comment is considered a "document". This technique allowed us to get samples of sentences from comments that most represented a given topic. We used NLTK and the spaCy English model to perform text pre-processing such as lemmatization, stemming, tokenization and the removal of stopwords. Figure 1 reveals a growing concern about the virus based on news that had begun to leak from China. As China is often considered to be where the virus originated, many commenters referred to COVID-19 as simply the "Chinesevirus". During this time, there were also numerous comments surrounding the idea that the novel coronavirus was related to the impeachment of U.S. president Donald Trump. The discourse then shifted to the idea that "bats" and then "monkeys" were the cause of the virus, respectively. In the month of April 2020 (Figure 2), comments began discussing the ideas of businesses being "shut", having to "stay" home, and issues with "work." Discourse concerning having to wear (or finding) a "mask" continues, as does the idea of the virus being a "lie" and animosity against the "news" media and "china" remains dominant. In the last part of April 2020, discussions concerning toxicity about "trump" were dominant.

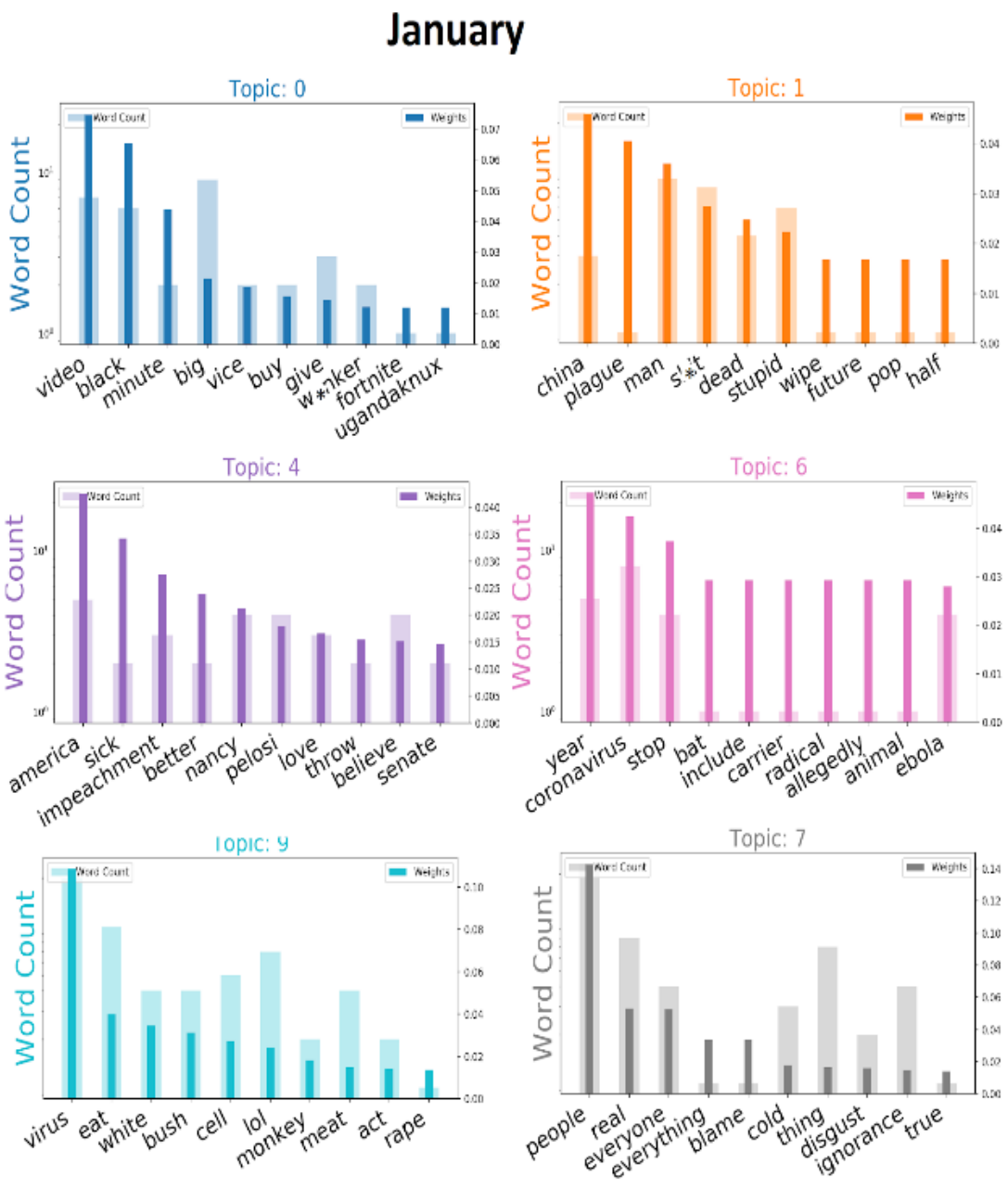

Figure 1. Word count and importance weight of the prominent topics for January 2020

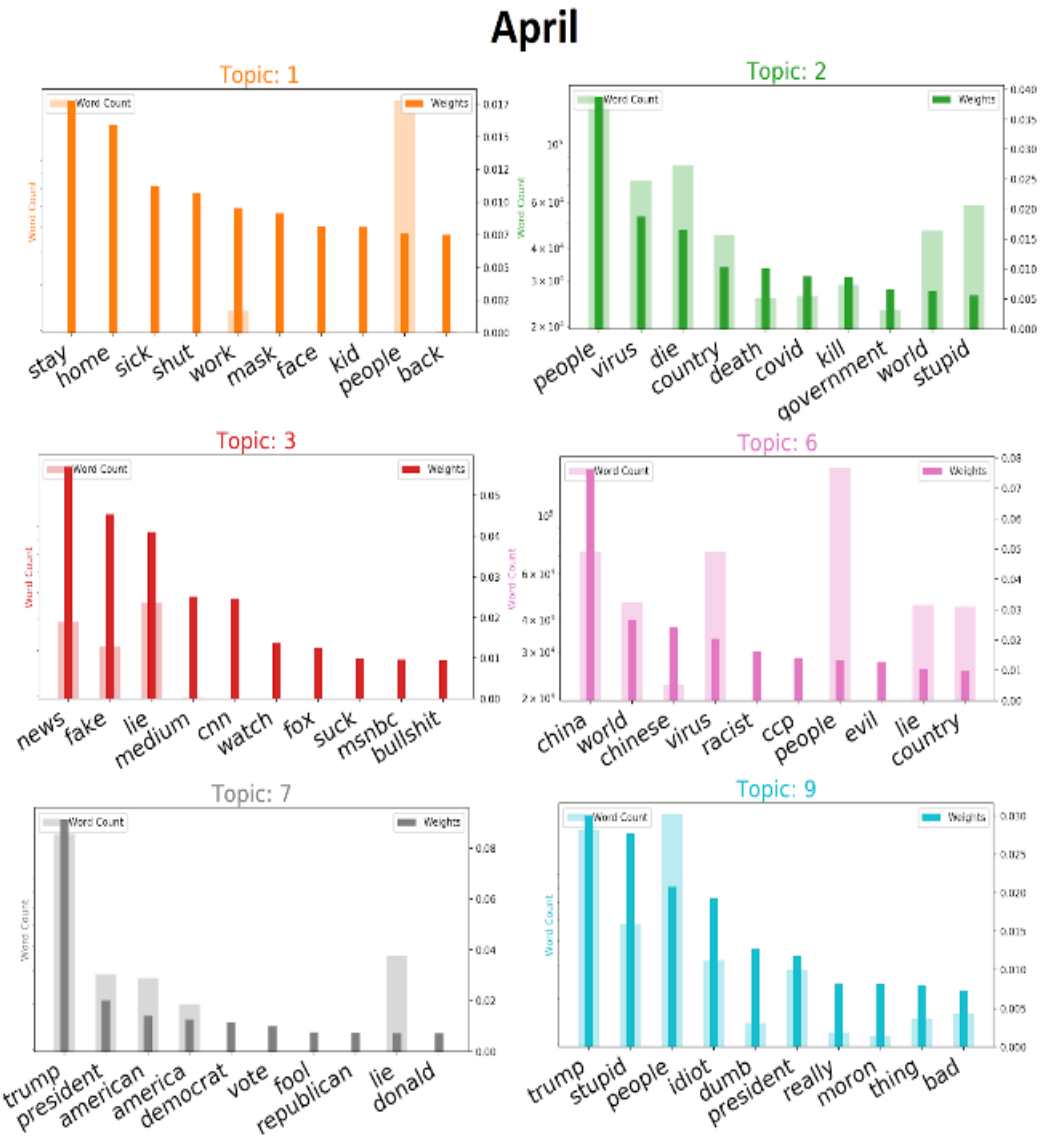

Figure 2. Word count and importance weight of the prominent topics for April 2020.

Social Network Analysis

We sought to analyze networks that profile common behavior patterns among toxic users [9], including leaving comments on multiple videos within a channel, replying to other commenters with toxic content, and repeating/duplicating comments across videos. The results of our analysis are detailed in this section. Social network analysis (SNA) was performed on 145 videos that were posted in110 YouTube channels, with a focus on the 32,107 unique users that left comments on these videos. In the accompanying network visualization, the nodes are colored based on their toxicity scores (0.5 to 1) with the color ranging from blue (lowest toxicity) to red (highest toxicity). Where appropriate, we discuss important centrality measures. SNA allows us to identify the top toxic users in our network, which helped to inform experiments simulating the impact of the removal of these users.

Co-Commenter Network

Toxic comments do not often exist in isolation. Instead, some videos attract multiple toxic users .

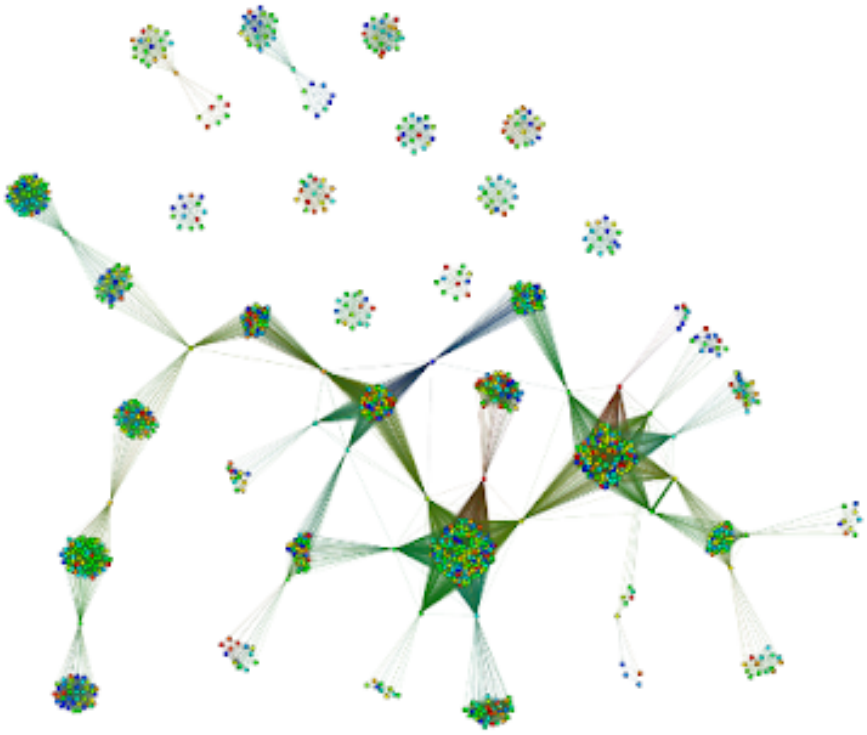

Figure 3. Co-commenter network. The nodes are commenters and the edges are shared videos.

Our Co-Commenter network analysis sought to examine the extent to which the users in our dataset commented on the same videos. We found that 2,453 users commented on the same video with at least one other user. In the network depicted in Figure 3, each node represents a user, and the edge represents a video in common. We modified the network by removing components such as dyads, triads and any cluster with ten or less commenters. After the modifications, 2263 users remained. The resulting network shows the biggest com-ponents within the network. We noticed that the largest connected component (2,057 nodes with 638,744 edges) forms a series of smaller clusters that share one common commenter. These commenters form bridges between the smaller clusters that enables them to spread information quickly. These users also act as gatekeepers, who have the following roles: selecting, channeling, shaping, manipulating, and deleting information. If any of these commenters are removed, the network structure will disintegrate. These users are also very toxic in nature.

Co-Commenter Shared Comment Network

The final toxic behavior pattern we studied was the tendency of toxic users to duplicate and repeat the same comment on multiple videos. The co-commenter shared comment network highlights this behavior. We found 117 duplicate comments shared among 213commenters. These shared comments consist of the exact text that multiple users posted which could suggest suspicious behavior. We modified the network by removing components such as dyads, triads and clusters with less than 5 commenters. These filters eliminated discussions from January 2020 completely, as the behavior was more common during the remaining months represented in our dataset (February to April 2020). The results are highlighted in Figure 4, along with some of the most repeated comments. Toxic users formed groups accusing the President of the US as “Racist”, “Idiot”, “Moron”, etc., while a few highly toxic users protested against China with hashtags such as “#chinesevirus”. The overall discussion focused on the novel coronavirus but there

also existed non-relevant content pushed by certain commenters such as "f***liverpool", "Russian troll farm conspiracy theories…", etc.

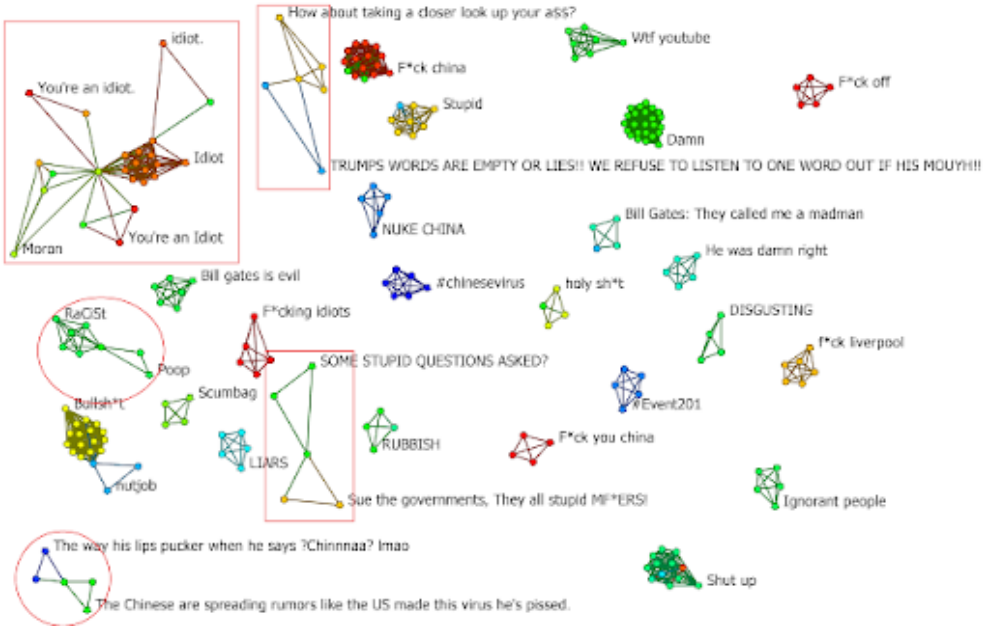

Figure 4. Co-commenter Network based on shared comments from February - April 2020.

Commenter-Reply Network

Toxic content is also often present as replies to video comments. Figure 5 shows our Commenter-Reply network, where we profiled the patterns among the 7.885 replies to comments present in our dataset. This was created as a directed graph to demonstrate the flow of conversation. Each edge represents a reply relation between a commenter and the replier with the arrow pointing towards the replier. We modified the network by removing components such as dyads, triads, and any channel cluster with less than 30commenters in order to make the network legible and focus on the larger clusters within the network. We found that the replies to toxic comments had a similar level of toxicity.

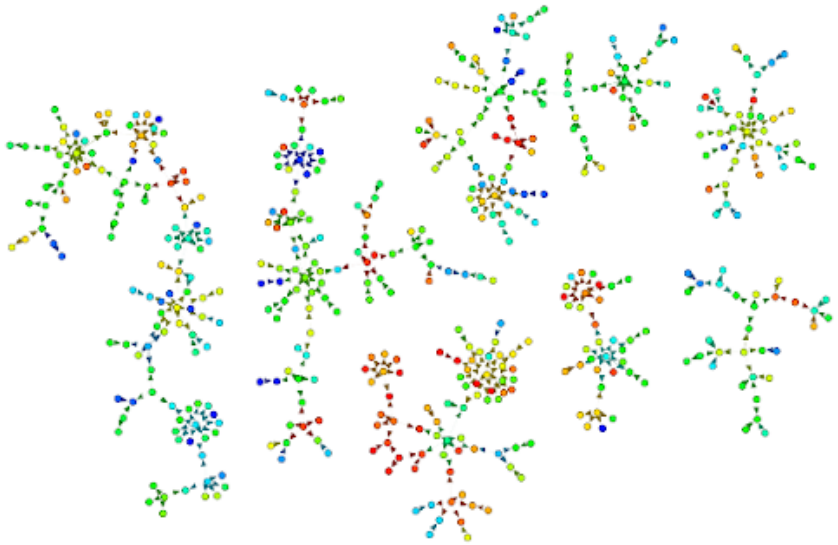

Figure 5. Commenter-reply directed network

Conclusions and Future Work

In this work, we collected data about YouTube videos related to the COVID-19 pandemic and analyzed the patterns within each video's comments. Through the use of topic modeling and social network analysis, we have detailed methods to 1) identify toxic comments on OSNs, 2) identify the common topics of those toxic discussions, 3) identify the commenters who are propagating that toxicity across a social network. Once we were able to understand the most toxic offenders and their behavior patterns, we were able to perform simulations to envision a network without them. Methods such as the ones utilized in this work can be useful when incorporated into the moderation processes of OSNs. When toxic commenters and their behavior patterns are identified, administrators of those social media platforms can decide to either flag or remove such commenters from the network, which would improve the overall health of the communication platform by reducing its average toxicity. This technique can be applied to any social media platform. The issue remains, however, of that fine line between censorship and trampling the users' right to speech. It will be up to the administrators of these various OSNs to decide. Future work will include an expanded set of search terms to capture a wider range of discussions, and an extension of our analysis to conversations that have occurred since April 2020, the end of our current dataset. Additional methods can be employed to discover patterns within the comments as well as among the behavior of each commenter. As the focus of this study was on YouTube activity, our work can be further validated by investigating similar behavior on other OSN platforms.

Acknowledgements

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 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), U.S. Air Force Research Lab, U.S. Army Research Office (W911NF-17-S-0002, W911NF-16-1-0189), 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. Obadimu, A., Mead, E., Hussain, M. N., Agarwal, N. (2019, July). Identifying Toxicity Within YouTube Video Comment. In International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation (pp. 214-223). Springer, Cham.
2. Varjas, K., Talley, J., Meyers, J., Parris, L., Cutts, H.: High School Students' Perceptions of Motivations for Cyberbullying: An Exploratory Study. West J Emerg Med. 11, 269–273 (2010).
3. Shachaf, P., Hara, N.: Beyond vandalism: Wikipedia trolls. Journal of Information Science. 36, 357–370 (2010).4
4. Lee, S. H., & Kim, H. W. (2015). Why people post benevolent and malicious comments online. Communications of the ACM, 58(11), 74-79.
5. Chen, Y., Zhou, Y., Zhu, S., Xu, H.: Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing. pp. 71–80. IEEE, Amsterdam, Netherlands (2012).
6. Warner, W., Hirschberg, J.: Detecting Hate Speech on the World Wide Web. In: Proceedings of the Second Workshop on Language in Social Media. pp. 19–26. Association for Computational Linguistics, Montreal, Canada (2012).
7. Sood, S. O., Antin, J., & Churchill, E. (2012, March). Using crowdsourcing to im-prove profanity detection. In 2012 AAAI Spring Symposium Series.
8. Martens, M., Shen, S., Iosup, A., Kuipers, F.: Toxicity detection in multiplayer online games. In: 2015 International Workshop on Network and Systems Support for Games (NetGames). pp. 1–6. IEEE, Zagreb, Croatia (2015).
9. Cheng, J., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2015, April). Antisocial behavior in online discussion communities. In Ninth International AAAI Conference on Web and Social Media.

10. Wulczyn, E., Thain, N., Dixon, L.: Ex Machina: Personal Attacks Seen at Scale.arXiv:1610.08914 [cs]. (2016).
11. Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., Leskovec, J.: Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17. pp. 1217–1230. ACM Press, Portland, Oregon, USA (2017).
12. Thomas Davidson, Dana Warmsley, Michael Macy, Ingmar Weber: Automated Hate Speech Detection and the Problem of Offensive Language. In: Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM2017).
13. Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving google's perspective api built for detecting toxic comments. arXiv preprint arXiv:1702.08138.
14. Grondahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018, January). All You Need is" Love" Evading Hate Speech Detection. In Proceedings of the 11thACM Workshop on Artificial Intelligence and Security (pp. 2-12).
15. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.