Managing Misinformation on Social Media Through Semi-Automated Methods

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Nicholas O'Connor

Spring 2022

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

S. Travis Elliott, Department of Engineering and Society

Background

Over the course of recent history, there have been many technological developments that improve the speed and efficiency of communication. These include the Telegraph telephone and Internet connection between personal computers as well as a larger support structure of postal services and intra organizational communication. But none have had so much of an impact on personal communication as the development of social media networks. Not only do they provide near instantaneous communication as with the development of the Internet, they also provide means of accessing contacts faster than could traditionally be acquired (such as through getting someone's phone number or instant messaging address) through recommendations of contacts that a user is likely to have already known.

These developments in communication technology alone are already impressive in the speed at which new communication routes can be formed. However, those alone do not account for the omnipresent access to and use of online social networks. Developing in parallel with these more powerful forms of communication, new devices became available to consumers which allowed them to be much more flexible with how they communicated. It started with home computers that gave the average person access to the Internet and social networking, which then improved with faster and more broadly connected Internet to reach more and more possible contacts. Then, with the development of portable and affordable smart phones, such as the development of the iPhone, users now have the ability to use these methods of communication from anywhere; not just at their home or place of business, where they were limited to only having access to a PC or a telephone landline.

With devices becoming more accessible, more portable, and more powerful, users started being able to generate and consume media at an exponentially increasing rate. Access to cameras

in any location allowed more pictures to be taken and more videos to be shared of current activities. More powerful media creation tools allowed for generation of media at a faster rate. Faster Internet connections allowed for more media to be shared in a shorter time window. By current estimates, there will be 463 exabytes (463,000,000 terabytes) of data being created and shared on the Internet per day by 2025 (Vish, 2020). Observing and going through all of this information to find media that is relevant to a specific user is nigh impossible. A solution was needed so that users would be able to find media that was relevant to their interests.

To solve this open market, social media companies created web applications that allowed users to find content they would be interested in. Some of the first iterations of social media, such as Myspace, had the user curate their own experience by specifying which friends they wanted to see status updates from. However, this still had the previous problem of a user needing to go through information on Myspace to find a friend's account or get that contact information from the person themselves. With the development of Facebook hey feature was supplied to users so that they could find their friends, and their friends shared content, without needing to contact that friend directly. Contacts that were likely to have been known to a user would be suggested to that user by Facebook so that restart of getting the address of contact could be removed from the process of expanding a social network.

Recent Examples

To give some examples of how the rapid spread of information with no checks as to its veracity, I first examine a recent controversy from 2017 named PizzaGate. Who exactly started the rumors behind PizzaGate are unknown but sources point towards posts on Facebook that started circulating regarding a pedophilia ring involving prominent American lawmakers such as Hillary Clinton. As these rumors continued circulating, online news sources began to pick up the

rumors and create articles about how people were talking about these rumors involving high profile lawmakers and pedophilia claims. These news sources were then taken as sources of fact, which led to their broadcast on more widely viewed news media such as Infowars hosted by Alex Jones. The consequence of a rumor percolating up from one anonymous source to a widely viewed broadcasting source led to the consequence of an armed man entering Comet pizza in Washington DC, to free children from a pedophilia ring in the restaurant's basement, of which the basement did not exist.

The above examination of PizzaGate shows how one, possibly unintentional, source of misinformation can spread and amplify to cause more tangible consequences. However more recent cases have shown that a small group of coordinated individuals can synchronize their actions to artificially amplify the effect of seeds of rumors into a larger movement. During the 2020 U.S. presidential election, a large campaign was conducted that claims that the results of the election were fraudulent based on many unverified sources. Resources wood purport to provide video or written evidence that specific ballots are being overturned in contentious polling districts, which would then be picked up by higher profile public figures who would then report it as fact. The sources that made these claims were not necessarily acting independently to make these claims, but would often work with the public figures to appear that they were just picking up on word-of-mouth claims when in actuality they were supplying the evidence themselves. The speed at which these claims would travel, as well as their sensationalist descriptions, made it difficult to both fact-check in the first place and also do so quickly to stem the travel of the original claim.

STS Framework

The framework used for this analysis will be the social construction of technology (SCOT). SCOT claims, as a theory, that technology alone does not shape human action. Rather, human actions and behavior shapes how technology is developed and applied to problems. There are free fields to consider when applying SCOT to any sort of technology. First, one must look at the relevant social groups. Different groups of people have different interpretations of the problem as well as different interpretations of what a solution to the problem should actually solve. Second, there is the design flexibility that should be considered. The discussion over the solution should not be over whether one design is the correct solution or should not be implemented at all, rather it should be acknowledged that there is more than just one proposed solution available within the design space. There are many different tweaks and possibilities as well as entire reconstructions of a solution throughout all of the design space available. Third, one should consider the problems and conflicts involved in both the initial problem as well as any solutions to the problem. Consider what conflicts led to the creation of the problem in the first place as well as any problems that may arise from the implementation of a given solution or any solution in general.

We can now use SCOT to apply it to the problem of moderating and managing misinformation and aggressive behavior on social media. First, we start with relevant social groups. These include the users of the platform, of which there are users who do and do not intend to use social media for gaming or spreading information. There are also the business owners of social media applications, who seek to ensure that they make money off of whatever business model is supported by the application. And finally, there are people who do not interact or use social media applications directly but have some connection to people who do. Of these groups, there can be a case made that all of them have a stake in reducing the amount of

misinformation on a social media platform, save for people who wish to use said platform to intentionally spread misinformation. Users of the platform who did not wish to use it for gaining or spreading information would likely have a general concern that the application does not contain this information in general, all those who do use it to gain and spread information would be more strongly concerned that there is less misinformation on the platform, but also that they can trust that their own information that they believe to be true and sincere is not stopped by procedures that are intended to stop misinformation.

We can also consider design flexibility in regards to solving this problem. There are not just solutions that exist within the current structure of a social media design, but also ultimate designs for social media that allow for a broader range of possible solutions. Finally, we have the problems and conflicts that arise that cannot be solved technologically such as how information is not spread solely through social media but also through offline connections between personal relationships after information has been received through social media. There is also the consideration of legality and how any proposed solution interacts with the laws of a specific jurisdiction in particular the general consensus of the right to freedom of speech.

Current Methods of Moderation

Facebook

Facebook is the most popular social media network currently in use. Its features include a wide range of methods of connecting to different people, communities, and organizations.

Facebook's algorithm does not just suggest content that it thinks the user would be interested in, it mainly suggests content that it thinks will keep the user on Facebook for as long as possible. While this usually intersects with what a user is interested in in most cases this content will cause a user to experience increased emotional reactions as well as a higher

likelihood to share the content with their friends. Unless careful curation of a social media feed is performed this usually leads to content that is mainly sensationalist exaggerated and intended to raise emotions such that the user feels compelled to take action. This typically results in sharing this information with their friends in order to encourage them to also take action, or more time spent on Facebook learning more information. As misinformation is typically by nature also sensationalist and intended to drive emotion, it is given higher priority by the content recommendation algorithm.

If this misinformation does occur on Facebook, there are a few methods that Facebook uses in order to curb its efficacy. Part of its process can involve user reports, where a user can flag a post for review by Facebook moderators. However, because so much content is created and shared on Facebook per second, it is unfeasible to manually review and take action on every single user submitted report. Therefore, Facebook also uses automated methods to find keywords in the content of a post to determine if that post is likely to contain primarily aggressive or harmful content to the site in general. It is unable to check if the content of a post is strictly untrue primarily because it is currently near impossible to determine that using purely automated processes.

The last option that Facebook gives as a course for moderation is administrative control over a Facebook community: where any post made within the community is posted to all members who follow the community's content. In this Facebook community, they are given the ability to delete posts or ban members from the community they have created. Aside from the stipulation that all posts within the community must comply with sitewide guidelines, there is no obligation as to what the community is used for.

Twitter

Twitter's user model is based around conglomerating all content that a user wants to see into one unified feed. This is similar to how Facebook has one main feed for its user, but without the flexibility of groups and communities. In order for a Twitter user to add content they wish to see to their own feed, they "follow" the account representing the person, celebrity, or organization they want to see the content from. Each "Tweet" from that account will then appear on their own feed, alongside all other content that user has followed. Twitter has also recently released an alternative feed style. It aims to mimic how Facebook provides content to it's users by allowing an algorithm to select content for the feed in a way that maximizes user engagement. Instead of only the tweets of accounts that the user is following, the user's feed is composed primarily that category, as well as tweets that were simply liked by the followed accounts, as well as a higher density of suggested accounts to follow. These tweets are also not necessarily presented in a chronological manner, as opposed to the standard feed style, which enables Twitter to provide a theoretically infinite amount of content, until the user decides they've had enough. However, users seem to prefer the original, chronological feed method, despite Twitter attempting to promote the algorithm-curated one.(Newberry & Sehl, 2021)

Thus, much of what a user sees using Twitter is what they would expect to see from personalities they reasonably understand and enjoy interacting with, save for the re-tweets that could reach them from a friend of a friend of a friend. This causes communities to form around shared interests, often with a user existing in multiple circles at once. (Thompson, 2019)

According to their transparency data, more than 4.8 million accounts have had come form of disciplinary action applied to them between January and June 2021 alone (Twitter, 2022). At a rate of approximately 26,000 actions per day, this is clearly more than can be handled by Twitter's 5,000 employees, many of whom are focused on matters other than moderation

(Statistia Research Department, 2021). However, I have not been able to find exact specifications on how Twitter automates their moderation activities, only sources on how user-submitted reports are handled, reviewed, and enforcements applied. These enforcements include putting notices over tweets suspected to contain harmful content—such as tweets casting doubt on the results of the 2020 presidential election (Parrott, 2020)—deleting offending tweets, or up to permanently suspending an account entirely. An appeals process is available, should a user dispute the reasoning behind a disciplinary action.

Reddit

The third platform for social media that will be examined is Reddit. Reddit has a drastically different method of use when compared to Facebook and Twitter. Content is sorted into different "subreddits", such that each subreddit is focused on a specific topic. A subreddit's title will often be prefixed by "r/" in order to reflect how it appears in the site URL. The definition of what topic can be chosen for a subreddit is very broad, ranging from the very general "r/news", where users post and discuss news articles regarding current events, to the very specific "r/breadstapledtotrees", where users share pictures of one subject only: bread stapled to trees. (Boyd, 2022)

Like Facebook and Twitter, any post to Reddit as a whole is subject to their site-wide rules. These share similar sensibilities to the former two, such as disallowing calls to violence, but also including rules that pertain specifically to Reddit's structure, such as disallowing the members of one subreddit brigading the posts of another. Much more responsibility is given to a subreddit moderator, as opposed to an administrator who oversees a Facebook group, in regards to enforcing both subreddit and site-wide rules.

Conclusion

If one person in a friend group is being aggressive, it's very easy for someone else in the current conversation to step up and call out that person for whatever aggressive comments they might be making. That then leads to other people in the same conversation typically giving their own shame on the person who has spoken those aggressive comments. After that, there isn't much for that person to continue doing in that space. Clearly, they could still talk with those people, but without reneging on what they had already said, or if they want to continue talking about whatever they had said, they have to go to some other space. This natural flow of authority to someone challenging the negative behavior of someone else is what manages negative behavior in physical spaces. In online spaces, this isn't as possible. The reach that a malicious actor can have in an online space is orders of magnitude larger than what is possible in a physical space. Additionally, there isn't an available call to fluid authority such that someone can be cast out of a conversation. Typically, the only accepted authority in an online space to handle conversation that is either off topic, malicious, or insensitive is either not always available in case of a smaller number of moderators that are not always online all the time, or not always trusted in the case of moderators of larger platforms of whom the authority is not entirely accepted by who use the platform.

For this reason, the structure provided by Reddit's model seems to be the closest to how social management of misinformation happens in real life. Smaller groups, focused around a specific topic, have a smaller team of people who have been given an authority to manage the conversation. While there is some mistrust in the fact that there is not consideration given to the approval or disapproval moderators from other users of a certain subreddit, authority seems to be accepted based on the fact that the scope of the authority is limited only to that topic as opposed

to sitewide administrators having to manage every possible conversation. Obviously, not all solutions can be solved by this model. However, this leads to the implication that the problem of dealing with misinformation and aggressive behavior on social media cannot be solved alone by stronger moderation teams within a social media company, but instead requires a stronger look at the structure behind how a social media network is used by its users. A conversation must take place the developers and engineers and business lines behind a social network idea, and the potential people who will use the social network, in order to fully capture how to more closely mimic how actual socialization behavior performs in physical spaces. Mapping structures onto human behavior is more effective than attempting to force human behavior to fit within an artificial structure.

Works Cited

Boyd, J. (2022). *What is Reddit?* Brandwatch. https://www.brandwatch.com/blog/what-is-reddit-guide/

Newberry, C., & Sehl, K. (2021, October 26). How the Twitter Algorithm Works in 2022 and How to Make it Work for You. *Social Media Marketing & Management Dashboard*. https://blog.hootsuite.com/twitter-algorithm/

Parrott, J. (2020, November 6). *Twitter suspending fake news accounts, labeling tweets that peddle 2020 election misinformation*. Deseret News. https://www.deseret.com/indepth/2020/11/6/21552939/twitter-election-2020-fake-news-associated-press-trump-biden-dorsey

Statistia Research Department. (2021, April 1). *Number of Twitter employees 2020*. Statista. https://www.statista.com/statistics/272140/employees-of-twitter/

Thompson, R. (2019, March 25). *Twitter cliques might feel like high school, but their existence is tied to our human nature*. Mashable. https://mashable.com/article/twitter-cliques-high-school-human-nature

Twitter. (2022). *Rules Enforcement—Twitter Transparency Center*. https://transparency.twitter.com/en/reports/rules-enforcement.html

Vish. (2020, June 24). *How Much Data Is Created Every Day in 2021? [You'll be shocked!]*. TechJury. https://techjury.net/blog/how-much-data-is-created-every-day/