# The Growth in Misinformation Over Social Media Networks and the Change in Content Moderation Over Time

A Research Paper
in STS 4600
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Vraj Desai

April 15, 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

Signed: Vraj Desai Date: 04/15/2022

Approved: Kent Wayland Date: 05/13/2022

Kent Wayland, Department of Engineering & Society

## Introduction

Currently, the global pandemic caused by COVID-19 has had 245 million reported cases, and 4.9 million deaths, which has put the entire world in disarray (International Telecommunication Union, 2021). Furthermore, the quick growth of the Internet social media technologies during the pandemic has dramatically changed school life, jobs, social lives, and much more. In such an environment, multiple studies conducted have shown an increase in the spread of misinformation over social media. COVID-19 misinformation may mislead users and negatively impact society's culture, economics, and healthcare. Additionally, the spread of misinformation related to COVID-19 and its vaccination could lead to many obstacles and affect health care workers poorly. Furthermore, many of these large social media platforms have taken Human Content moderation efforts and newer machine learning approaches to prevent the spread of such misinformation. The completed STS research serves as an accessible primer on how content moderation is conducted by large social media platforms. It will analyze existing automated systems used and emphasize some of the larger issues many of these machine learning systems include as they are developed and utilized. Findings from this paper will provide insight into a proper approach that can be taken by large corporations to prevent the spread of COVID-19 misinformation. Many companies take upon different approaches to content moderation. For instance, Facebook does not flag subjective content as fake but rather flags it as misleading. Any sort of unintentional misinformation or disinformation flags it as fake news. On the other hand, Twitter uses a more aggressive approach, where it flags tweets as harmful, and these can be tweets that particularly cause physical, psychological, or informational damage.

One of the most recent and common methods of misinformation detection is a machine learning model that uses natural language processing classifiers such as LSTM, RNN, and KNN.

These tools can easily find the difference between factual and nonfactual pieces of information. (Al-Smadi, M., Qawasmeh, O., Al-Ayyoub, M., Jararweh, Y., Gupta, B, 2017). Since misinformation will spread rampantly, multiple machine learning algorithms are currently being developed to combat this issue.

Many machine learning models will not reach full accuracy and also have to test multiple models to see if it really is detecting misinformation. Hence, there will be a large cost that will go into developing various models and then testing their effectiveness. Furthermore, many companies are not taking action against fake news threads and misinformation proliferation. Costs may also include hiring software developers to create machine learning models that are able to immediately detect widespread misinformation and remove it or ask the company to investigate. Additionally, there are numerous challenges with commercial content moderation as enacted by popular platforms today. Costs tend to include labor concerns (poor working conditions and mental health challenges faced by moderators), democratic legitimacy concerns since corporation executives are setting the rules on free speech rules, and concerns regarding the overall lack of transparency and accountability (Gillespie, 2018).

Consequently, individuals may be surrounded by non-credible sources, and once people become fixated on an ideology, it is extremely difficult to alter their viewpoint. Initially, this paper serves the purpose of emphasizing the magnitude of this issue and discussing the potential solutions. This paper hopes to provide a thorough view on content moderation and how different approaches have been successful and unsuccessful. As seen through the divide in the United States, the country is split on the best course of action to end this pandemic. People often cite this misinformation when proving their opinions and the validity of their actions. The consequential health effects of a mask, the ineffectiveness of the vaccine, and conspiracies about the origin of the coronavirus have all been cited through sources of misinformation. When important

information regarding this pandemic is not taken the correct way, the pandemic will only be prolonged.

This matter has become a large social issue because of the way social media has evolved. Since the 1970s misinformation has been widespread and caused false beliefs about many socio-political issues. In the 2016 presidential election, there were large amounts of misinformation being spread through Facebook and Twitter. Since the 2016 election, we have seen a massive increase in content moderation due to public attention and policy debates in government (Gorwa, 2019).

**Issues with Content moderation**

When it comes to identifying misinformation, there is a lot of backlash due to the different ways information is classified. This paper is meant to unpack the different types of misinformation and how different social media networks approach data and content on their sites. This can be misinformation that threatens public health, civic engagement, and election data. Hence, Twitter's approach has been extremely useful during the COVID-19 pandemic due to the spread of misinformation. In fact, if we look at efforts made outside the United States, we can see that the EU has taken many steps towards identifying misinformation by flagging any content that is defined as "false, inaccurate, or misleading information designed, presented, and promoted to intentionally cause public harm or for profit"(De Cock, 2018).

Due to poor actions taken by many social media companies, distrust can formulate amongst users. There is no actual uniform understanding as to what is problematic content – disinformation, misinformation, subjective data, "fake news." Sometimes a platform may only flag misinformation, but the users may be expecting a broader focus that also checks for disinformation. This creates distrust and often misleads many users because they may think all

"problematic" content is being flagged/removed. Consequently, they would believe such information, which is often misleading and this is often what happens with cases in the pandemic (Cock, 2018).

Recently in 2019, Facebook decided that it would reduce the number of groups and pages that promoted vaccine misinformation. It also promised that Instagram and other subsidiaries of facebook would also not advertise anything. However, reports discovered very quickly that anti-vaccination accounts on Facebook had a total of 58 million followers. (We have seen a rapid increase in social media companies taking efforts to ban certain users or takedown pages due to political and public pressure. Facebook has taken preventative measures such as signposting users to credible information sources, adding warning labels to misleading information, and also removing posts that could incite violence. Furthermore, they deal with content moderation through the removal of 3,000 accounts and 20 million pieces of stand-alone material (Ed Pertwee 2019).

**Literature Review:**

When looking at the literature review of "COVID-19–Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis" we see the different types of fake information that spread during the pandemic and how the situation has changed over the past few years. This literature aligns with the research paper as it shows the different types of false information. In such a study they were able to separate types of misinformation and put them into different categories to see how it all differs on portions of the internet. They were able to find countless amounts of misinformation and from which country they originated. This paper heavily relates to the research conducted in this paper as it shows the amount of misinformation

coming from one pandemic and we can see that growth compared to past occurrences. They use a different approach rather than the one utilized in this paper, but it still shows the growth of misinformation and the type of data that is spreading. Through this research we were able to realize that misinformation can be categorized into stigma, rumors, and conspiracy theories. Rumors regarding vaccine safety were dispersed via traditional media long before social media became the largest media outlet. There was much mistrust in society when diphtheria, pertussis, and tetanus vaccines were released during the 1970s and 1980s. Furthermore, this paper discusses the issue of rumors, stigma, and conspiracy theories when it comes to any sort of misinformation. A major issue with misinformation is the stigma and fear it creates in certain areas where there are strict belief systems. People tend to hide their symptoms because they are afraid of repercussions. Due to these types of misinformation we see a lack of trust among people in our institutions. It is very apparent that we have an immense lack of trust in our government, and as a result, we have an increase in anti-mask and anti-vax sentiment.

**Data Analysis**

Data that is collected regarding misinformation is compared with previous misinformation to see how bad the situation has become. We can now see the change in sentiment and trust between individuals due to the pandemic. We can analyze how the perception of the government's disease policies has changed over time. Furthermore, we can see the public reaction over social media regarding decisions made by public health officials. Many machine learning models conducted various exploratory studies over Twitter and Facebook and found numerous misinformation. One of the major studies conducted by Aswini Thota utilized a dense neural network approach for detecting fake news using FNC-1 data (Thota,

Aswini, 2018). The study began with data preprocessing and then their proposed model also identified the relationship between fake news and true news regardless of the relationship. When conducting their experiment, they concluded that their proposed model performed much better than any other researchers in the experiment. Evidently, the machine learning approach will not find all of the occurrences of misinformation. Much misinformation is deep-rooted by emotions within public sentiment. Further analysis will discuss how public sentiment has changed over time regarding the trust in our institutions. As a result, people form incorrect opinions and choose personal thoughts over actual logic.

Much misinformation and disinformation can be categorized and finding how those spread is the best way to approach this issue. When looking at the specific case of COVID-19, much of the data obtained will be reviewed and categorized into three categories: rumors, stigma, and conspiracy theories. Algorithmic detection is able to see the growth of these categories across the world through Facebook and Twitter. The amount of rumors has increased from February and continued to rise while peaking in March 2020. Many of these are related to COVID-19 sickness, transmission, and mortality. In terms of stigma, several healthcare workers were bullied or physically insulted by their landlords and peers if they contracted the virus. As the virus spread to various countries, stories went around that people of Asian origin were experiencing stigma (Frenkel S, Alba D, Zhong R, 2020). These instances are all meant to reinforce the idea that these forms of misinformation create distrust amongst users and hurt the foundation of our institutions.

**STS Theory**

When looking at the actors involved in a scenario like this, it is important to realize the

relationship between them. During the COVID-19 pandemic, multiple studies were conducted, which found that vaccine attitudes, trust in public health officials, and political beliefs were all interconnected. As a government, you have the duty to create trust between individuals and the institutions within the country. Trust can be seen as an abstract concept where "a relationship exists between individuals, between individuals and a system, in which one party accepts a vulnerable position, assuming the best interests and competence of the other, in exchange for a reduction in decision complexity"(Puri, N., Coomes, E. A., Haghbayan, H. & Gunaratne, K, 2020). This very idea has become an extremely relevant and important concept whenever there is a power imbalance or a transmission of information between two large parties. In the current situation of a public health issue, trust between all actors becomes apparent because there is so much social uncertainty. When it comes to the issue of trust in vaccines, masks, new public health guidelines, and the policy-making body itself. Distrust in our institutions and their subsidiaries is causing conflict and ultimately the spread of misinformation. As the world continues to carry forward with mistrust, social media only contributes to its spread and worsens the magnitude of the pandemic.

**Strategies**

There are a multitude of ways to dissolve many of the underlying issues that permeate with misinformation. There is one main method presented which is "increased transparency" (Abrams, 2021). And it can be implemented by building trust in users through transparency and taking measures that do not become censorship. Social media companies must cooperate with fact checking organizations, choose trustworthy algorithms that have broad coverage and limited bias, and link users to trustworthy sources more often with popular topics. In order to create trust

and increase transparency, platforms must make it clearer as to why a certain post is flagged. When Twitter removed a news article from the New York Post about hack materials, which was not allowed because Twitter's policy prevents content to be posted that is acquired by hacking. This was problematic because the reason was not completely apparent and Twitter received much backlash of removing content to protect its image. Hence, greater transparency amongst these issues in content moderation could make this problem easier. With a better form of transparency in who flagged the post and for what reason would create trust within the users (Abrams 2021). Specifically, it would also be beneficial for users to know if an item was removed due to an algorithmic or human content moderation. Early during the Covid-19 pandemic, Facebook has mistakenly blocked numerous posts from legitimate sources. The lack of transparency of who and why information was flagged exacerbated distrust and confusion in the users.

**Conclusion**

After looking through the various examples of content moderation and how they have changed over time, we have a better idea of the issues and benefits associated with it. With the COVID-19 pandemic these companies face the same struggles in appeasing their users in an informative manner. It is important for us to look at the best strategies to moderate content whether it is algorithmically or manually done. This paper pointed out many flaws within the current methods taken up by companies such as Facebook and Twitter. However, readers must glean on each of the examples and help devise a plan that ensures the safety of users while maintaining trust amongst them. When policy makers choose to address the issue of misinformation on social media and content moderation, they should consider the scope of their actions and how that will impact opinions politically, economically, and socially. Furthermore,

the nature of this issue is sensitive and requires much thought on how controlling the government or the social media sites want to be. If legislation was created that addressed the activities of US based social media companies, then imposing rules and regulations in other countries would be difficult. Consequently, whatever laws are passed or any decision that these large social media companies should be broad and transparent. This will ensure that there is limited backlash and the issue at hand is being addressed. From all of the given data and instances we can see how complex the issue of content moderation during the pandemic has become. In such times, we call upon our lawmakers and corporations to make sound changes to their policies and create a profound impact on our broken social media.

References

Abrams, Z. (2021, March 1). *Controlling the spread of misinformation*. Apa.org. https://www.apa.org/monitor/2021/03/controlling-misinformation

Al-Smadi, M., Qawasmeh, O., Al-Ayyoub, M., Jararweh, Y., Gupta, B. (2017, November 14). Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews. Journal of Computational Science. Retrieved November 1, 2021.

COVID-19: Science against fake news and misinformation. reddit. (n.d.). Retrieved October 24, 2021. Cox, K., Slapakova, L., & Marcellino, W. (2020, June 25). A machine learning approach could help counter disinformation. RAND Corporation. Retrieved October 25, 2021.

Cox, K., Slapakova, L., & Marcellino, W. (2020, June 25). A machine learning approach could help counter disinformation. RAND Corporation. Retrieved October 25, 2021.

de Beer, D., & Matthee, M. (2020, May 5). Approaches to identify fake news: A systematic literature review. Integrated Science in Digital Age 2020. Retrieved October 24, 2021.

De Cock Buning, M. (2018). A multi-dimensional approach to disinformation : report of the independent High level Group on fake news and online disinformation. In cadmus.eui.eu. Publications Office of the European Union. https://cadmus.eui.eu/handle/1814/70297

Demartini, Gianluca & Mizzaro, Stefano & Spina, Damiano. (2020). Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities

*DEPARTMENT OF EVIDENCE AND INTELLIGENCE FOR ACTION IN HEALTH*. (n.d.).

https://iris.paho.org/bitstream/handle/10665.2/52052/Factsheet-infodemic_eng.pdf

Frenkel S, Alba D, Zhong R, 2020. Surge of Virus Misinformation Stumps Facebook and Twitter. The New York Times. New York, NY: The New York Times.

Gillespie T (2018) Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media. New Haven, CT: Yale University Press.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 205395171989794. https://doi.org/10.1177/2053951719897945

Islam, M. S., Sarkar, T., Khan, S. H., Mostofa Kamal, A.-H., Hasan, S. M. M., Kabir, A., Yeasmin, D., Islam, M. A., Amin Chowdhury, K. I., Anwar, K. S., Chughtai, A. A., & Seale, H. (2020). COVID-19–Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis. *The American Journal of Tropical Medicine and Hygiene*, *103*(4). https://doi.org/10.4269/ajtmh.20-0812

Puri, N., Coomes, E. A., Haghbayan, H. & Gunaratne, K. Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases. Hum. Vaccines Immunotherapeutics 16, 2586–2593 (2020).

Social media: Misinformation and content ... - congress. (n.d.). Retrieved April 15, 2022, from https://crsreports.congress.gov/product/pdf/R/R46662

Stewart, E. (2021). Detecting Fake News: Two Problems for Content Moderation. *Philosophy & Technology*. https://doi.org/10.1007/s13347-021-00442-x

Thota, Aswini; Tilak, Priyanka; Ahluwalia, Simrat; and Lohia, Nibrat (2018) "Fake News Detection: A Deep Learning Approach," SMU Data Science Review: Vol. 1: No. 3, Article 10.