Using Natural Language Processing to Discern Misinformation in Online Media
(Technical Topic)

Defining Misinformation in a Meaningful and Actionable Way
(STS Topic)

A Thesis Project Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering

Nicholas O'Connor

Fall, 2020

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.

Signature _____

Approved _____ Date _____
Kathryn A. Neeley, Associate Professor of STS, Department of Engineering and Society

Introduction

The first major contender for the public space was MySpace, founded in 2003, with Facebook entering the scene less than a year later. In the nearly two decades since these applications first became available to users, as well as other services such as Twitter and Instagram, their ubiquity and reach has expanded far beyond any other innovation, becoming an almost required tool for anyone who has access to the Internet. In doing so, social networking sees an incredible amount of content being generated and shared every second, on the scale of four petabytes a day just on Facebook's servers (Vish, 2020)**.**

There has been a call for impetus behind research and development into the analysis and processing of this flow of data, as the nature of the unrestricted flow of information has led to *all* forms of information finding equal purchase within the public sphere, including information that is misleading, false, or malicious. The effect of such misinformation is especially notable after the events of January 6th, 2021, where the false idea that the results of the 2020 election were fraudulent were continuously spread within social networks, leading to the attack on the Capitol by people who deeply believed that they needed to take into their hands the course of the election certification in order to rectify the "incorrect" results. If misinformation continues to be disseminated in this manner, with no check on its progress, it is likely that these events will happen again. (United States Congress House Committee on Homeland Security, 2020, p. 24)

In this prospectus, I propose that a semi-autonomous system be developed in order to assist human moderators in identifying and removing intentional misinformation. This prospectus will also discuss how best to actually define such misinformation, as the range of definitions is broad and already a large subject of public debate.

Technical Topic: Using Natural Language Processing to Discern Misinformation Online

Machine learning (ML) is a branch of computer science that leverages a computer's ability to perform rapid calculations, and apply it in manners that allow it to exploit patterns that could normally only be spotted by human recognition. This is achieved through applications of linear algebra, where a model is trained on large amounts of data to adjust parameters until it can predict outcomes from future data with high confidence and accuracy. For this technical topic of the prospectus, we are using a technique of ML called natural language processing (NLP). This is a method of Machine Learning that enables algorithms to understand the intent and meaning behind human conversation. (Kosolwattana, Tanapol, 2020, p. 2)

The ability to train computer models to analyze the intricacies of human communication is important due to the inability of humans to keep up with the massive amount of data flowing through the internet, and social media networks in particular. An automated system is needed to help keep up with this flow of information. While the process of constructing and training a Machine Learning model on data does take some time, once completed, it can deliver results at high speed while wading through terabytes or even petabytes of data at a time.

Currently, social media companies already spend significant resources to moderate their platforms for general nuisance behavior. Even smaller communities, such as groups within Facebook's groups or subreddits around a topic on Reddit, have difficulty managing trolls, harassment, and misinformation. Much time is spent wading through meaningless data, hoping to pick out the offenders, without the ability to keep track of repeat offenders. Additionally, it is difficult to match patterns of behavior to notice malicious users that spread their activity over multiple accounts. Tracking down misinformation is even more difficult, as users may be spreading it with benign intentions, having simply learned it from another source without

checking its correctness. The source of such false information is often hidden behind many links shared between users, and it can be impossible to track when it jumps offline to travel through word of mouth.

There are also systems that have recently been promoted by Twitter that alert the user if they may be entering into a hostile conversation when replying to another user's post on their feed (Peters, 2021). It is unclear how Twitter determines the likelihood of a fight, but it is likely they use a form of sentiment analysis. This is another field of study that seeks to analyze opinions and attitudes This model would build on applications such as Twitter's alert system, both literally by using pre-trained data from currently available sentiment analysis models in a process called transfer learning, and figuratively by using the lessons learned from the construction of sentiment analysis models to flesh out the structure of a model to identify misinformation.

STS Topic: Defining and Sourcing Misinformation in Online Networks

It is difficult to comprehend the massive scale that social media networks encompass. Even if one says that "Facebook has 2.8 billion monthly users" (*Facebook MAU Worldwide 2021*), it's impossible to put that number into perspective in a way that a person can understand. A conclusion that can easily be made from this observation is that there is no way for a small group of people to keep track of all the information being shared among users. As discussed above, it is important to develop a system to identify misinformation so that action can be taken to slow it's influence. However, what could be classified as "misinformation" is tricky to resolve. Typically, misinformation is defined as statements given with an intent to mislead, but it is not always the case that the person spreading the misinformation did so knowingly, possibly thinking

that it were true instead. Understanding how such situations arise, and how the information that causes them spreads is important to furthering an understanding of how exactly to stop it.

In regards to social media posts which involve misinformation, it is challenging to track exactly where the misinformation came from. The easiest way to find the source of a piece of information or opinion is usually by a website's share feature, which is what easily allows media to spread as users share it with others who follow them. Social media posts are presented as from the user who originally created it, with a note as to who shared it so that it appeared on your feed. With this site functionality, it's clear to see where some source came from. The link is not as clear when the path of information is not contained within a social media service, or even within the internet entirely. Of course, communication can still happen offline through word of mouth. A user may see a post, remember its message, talk to a friend about it, and then that friend may create a post to share the information without a way to link to where it came from. Or, a user may share the information in a post in a transformative manner. Information sharing that is not fully connected is common with memes, where the message is communicated through a metaphorical medium that is already recognized by the audience with whom the user shares their post.

An example of a meme being used to communicate by relation is shown in Figure 1 (DHMO Awareness, 2020). The message it aims to convey is a dismissive view of many anti-vaccine arguments, by forming a mocking rendition of the form of argument

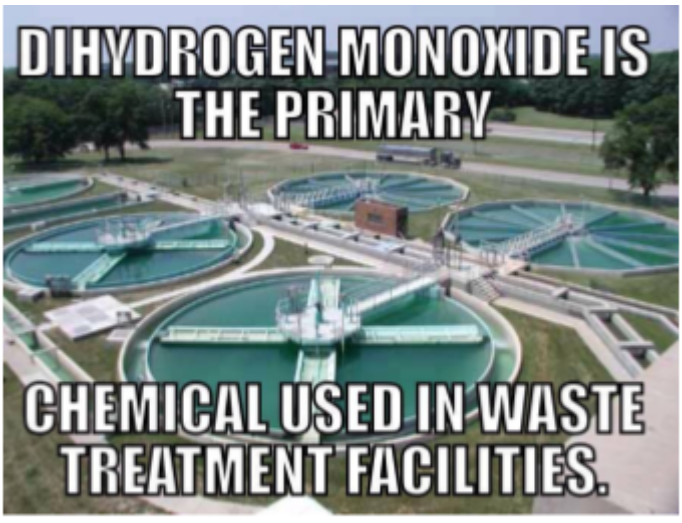

Fig 1: A satirical meme that mimics arguments used against various chemicals (credit @DihydrogenAware on Twitter)

typically used by members of that view, and changing the perspective to show absurdity. Notably, there are many layers to the message that is being conveyed within this image, which obfuscates the original source even more. All of these convoluting factors make it difficult to find the original source of any information. It even makes it harder to determine if the original source was intentionally sharing deliberate misinformation. There is always the possibility that there was some offline link behind what the user chose to share. (Gallo et al., 2021, p. 5)

There are many reasons as to why it is so important to stop misinformation, intentional or not, quickly. False information tends to be inflammatory, or contain a message which readers would typically react with intense emotion. The reactions that readers experience leads them to feel as though they need to take action quickly, and part of this action is often to inform their network of the information so they can commiserate on it. In contrast, information that is factual tends to be more detailed, and requires more thought to fully understand and internalize. In the environment of social media where the application is designed to constantly provide content, it becomes more difficult to engage with media that requires a longer attention span when another piece of media is ready to follow it up. (3Boxmedia (Firm), 2019)

There is also socio technical research already conducted that relates to the problem at hand. A paper authored by Amani Vohra (2020) performed an analysis of how search engines magnified the impact of misinformation with respect to medical practice, and how it changed patients' trust and behavior towards doctors. In particular, in its analysis of 673 COVID-related tweets, "The study also found that 24.8% of the tweets contained misinformation, and 17.4% contained unverifiable information." It is clear here that misinformation and unverifiable information is not an insignificant portion of the public discussion. (Vohra & Jacques, 2020, p. 17)

Conclusion

With the work completed from both the technical and STS research halves of this prospectus, we have two anticipated deliverables. On the technical side, we will have a robust model that is able to differentiate misinformation from factual information with high confidence. (Roberts, 2019, ch. 2 p. 38) From the STS research, we will have a better understanding of how misinformation is spread in social media and why it is difficult to combat. (Zhu et al., 2020, p. 5) On successful completion, these would provide the means for social media companies, and the communities and users that utilize those networks, to better manage the spread of misinformation and enjoy a safer experience online.

Word count: 2137

Works Cited

3Boxmedia (Firm). (2019). *Battle of social networks* (Internet materials). 3Boxmedia. http://proxy01.its.virginia.edu/login?url=https://fod.infobase.com/PortalPlaylists.aspx?wID=98131&xtid=189751

DHMO Awareness. (2020, December 8). Yuck! Https://t.co/8VgHsbIfyy [Tweet]. *@DihydrogenAware*. https://twitter.com/DihydrogenAware/status/1336108666928451586

Facebook MAU worldwide 2021. (n.d.). Statista. Retrieved October 25, 2021, from https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Gallo, J. A., Cho, C. Y., & Library of Congress (issuing body). (2021). *Social media* (Internet materials; CRS Report for Congress, Misinformation and Content Moderation Issues for Congress). Congressional Research Service. https://purl.fdlp.gov/GPO/gpo157644

Kosolwattana, Tanapol. (2020). *Toxic Tweet Classification with Natural Language Processing and Machine Learning Techniques; Mediators for Game Streaming* [University of Virginia]. https://doi.org/10.18130/V3-3R04-TM69

Peters, J. (2021, October 6). *Twitter's latest pre-tweet prompts let you know when you're about to jump into a Twitter fight*. The Verge. https://www.theverge.com/2021/10/6/22713211/twitter-pre-tweet-prompt-fight-intense-conversation

Roberts, S. T. (Professor of information studies). (2019). *Behind the screen* (Internet materials). Yale University Press. http://proxy01.its.virginia.edu/login?url=https://www.jstor.org/stable/10.2307/j.ctvhrcz0v

United States Congress House Committee on Homeland Security. (2020). *Examining social media companies' efforts to counter on-line terror content and misinformation* (Internet materials; Hearing Before the Committee on Homeland Security, House of Representatives, One Hundred Sixteenth Congress, First Session, June 26, 2019). U.S. Government Publishing Office. https://purl.fdlp.gov/GPO/gpo131513

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/

Vohra, A., & Jacques, R. (advisor). (2020). *Synthesis of medical misinformation on search engines and social media before and during covid-19* [University of Virginia, School of Engineering and Applied Science, BS (Bachelor of Science), 2020]. https://doi.org/10.18130/v3-tfy8-8c86

Zhu, H., Shen, H. (advisor), & Ferguson, S. (advisor). (2020). *Deep learning based occupants'*

activity prediction in a smart building assistant system; how to combat fake news: An examination of information literacy education [University of Virginia, School of Engineering and Applied Science, BS (Bachelor of Science), 2020]. https://doi.org/10.18130/v3-j4n3-c406