Prospectus

Creating a Chrome Plugin that Flags Misinformation
(Technical Topic)

Social Media Algorithms and Radicalism
(STS Topic)

By

Ryan Robinson

March 23, 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: Ryan Robinson
Technical Advisor: _____
STS Advisor: _____

Introduction:

In the past two decades, the internet and social media have become powerful tools that have facilitated the spread of information in many new ways. While this has had many positive effects, unfortunately it has also allowed for the spread of misinformation. Ideas that before the internet would never had been spread have been able to grow and gain much larger audiences than ever before, creating many more opportunities for the average person to become misinformed or politically radicalized. On a large scale, this has many disastrous effects. Extreme political parties and figures have been able to gain more power and support, leading to more radical legislation (Rauchfleisch & Kaiser, 2020). Once niche conspiracies opposing lifesaving public health measures such as vaccines have been propagated and normalized, weakening our society's ability to respond to public health crises (Broniatowski et al., 2018). Some radical online groups have even been considered an entry point into even more extreme radicalization leading to acts of terror (Ribeiro et al., 2020). In my research, I intend to approach this problem from two different angles. In my technical paper, I plan on using machine learning to create an easy-to-use chrome plugin that will automatically flag social media posts, articles, or videos that contain misinformation, or "Fake News". I will need to work on both creating an accepted, objective algorithm that can successfully filter out misinformation, and designing an effective, easy to use interface that will make my tool intuitive and more likely to be adopted. Hopefully, my tool will be able to be used by regular people to combat misinformation.

In my STS paper, I will research how different social media algorithms allow for a "Radicalization Pipeline" and the spread of misinformation. I will then analyze the overall large-scale effects of these behaviors as well as the effects on a smaller, individual level.

Technical Topic:

Before the internet, most people who were misinformed were such because they didn't have access to information. In this day and age, people face the opposite problem. There is so much information readily available that it becomes very difficult to discern what is fact and fiction. This becomes even more difficult when you introduce bad actors who actively are trying to deceive you for their own personal gain, as described by Broniatowksi et al. (2018) when they analyze the effects of Russian bots and trolls on the online anti-vaccine movement. The prevalence of misinformation, whether intentional or not, has many far-reaching effects.

The biggest effect is an overall increase in political polarization. It has been shown that users across all social media sites are much more likely to share and interact with inflammatory articles than more balanced, neutral ones (Faris et al., 2019). False or misleading articles take advantage of this and are often inflammatory on purpose, exponentially increasing the amount of content that could provoke a reaction from the average user and cause them to drift more to one side or another. The effect is a divided society that is not as accepting of other viewpoints and hostile to discussion. Each side's antagonistic view of the other is constantly reinforced with an endless stream of provocative, often misleading content. Many who wish to avoid this become apathetic to the entire political process and stop participating, allowing more radical views to gain support where they otherwise wouldn't. While increasing awareness of so called "fake

news" will not singlehandedly end our current polarizing political atmosphere, pursuing this goal is a step in the right direction.

I plan to do this by creating a google chrome plugin that will be able to detect articles containing misinformation via a machine algorithm and flag them. The goal would be to have an intuitive, easy to use way of filtering out misleading information from users. This problem has two major technical challenges

The first and most obvious problem is the creation of the algorithm. Fortunately, "Fake News" detecting algorithms have been done before (Ahmad, 2020), and I have a lot of resources at my disposal to create mine. I will combine several of the discussed techniques when creating my algorithm. In addition, Rich Nguyen, my machine learning professor, will be an excellent resource should I need any advising.

The second problem is accessibility. The project needs to be easy to install and have an intuitive interface or it simply will not get used. Packaging the project in a chrome plugin will make it easy to install. For the interface, I will research the psychology behind general human computer interface design as well as google chrome specific interface design (Pérez, 2019). Any of the professors who teach Human Computer Interface classes will also be great resources regarding this topic.

The biggest obstacle for the success of my project is actually social. In order for it to have the desired effect, it needs to be considered objective and accepted by both sides. This will be incredibly difficult to achieve, but I plan to do my best to train my algorithm on as varied a dataset as possible to achieve maximum objectiveness.

STS Topic:

The first priority of social media platforms such as YouTube, Twitter, or Facebook is to get users to stay on them as long as possible by any means necessary. Each platform has developed their own algorithm to feed users whatever it determines will get them to stay on the longest. Unfortunately, feeding users more and more radical content is an effective way of getting them to stay on a particular platform. This has led to what many call the "Radicalization Pipeline," where a social media algorithm is incentivized to gradually recommend radical content to get users to engage on their platform, unwittingly putting them into a political echo chamber and influencing their views (Ribeiro et al., 2020).

One thing that makes these algorithmic pipelines so dangerous is that they don't just target people in the political sphere. Through intense trial and error, these algorithms have discovered that many specific online communities have a higher-than-average number of people in them that are susceptible to getting sucked into this type of content. The algorithms thus recommend more radical content to users who participate in these communities, ensnaring many people who otherwise would have had no interest in radical content.

To demonstrate this, some scholars have created software to map out the complex network of user interaction on sites like YouTube to visualize the general viewing path many online radicals go through. The map also allows us to see what communities the algorithms are originally targeting. For example, the Incel community, a misogynistic online group of men that has many far-right undertones, was shown to funnel vulnerable users from the gaming community, the conservative community, the community of users who consume content related to genetics and natural selection, and many more (Champion,

2021). Another similar network map found that German far-right users often started their journey watching meme videos, conspiracy videos, and once again gaming videos (Rauchfleisch & Kaiser, 2020).

This is clearly a problem, as people who without social media may be perfectly normal can become radical and dangerous to society. In an extreme example, Elliot Rodgers, a 22-year-old who on the outside appeared relatively normal, went on a shooting spree in 2014 killing 6 people. He attributed the attack to radical beliefs on sexuality that he gained from the online incel community (Branson-Potts & Winton, 2018). We know now that many modern acts of terror commonly start with the perpetrator becoming radicalized online (Ribeiro et al., 2020).

In addition to these incredibly extreme examples, the existence of these radicalization pipelines has wider negative effects on our society. Nefarious foreign actors can use information warfare and take advantage of these algorithmic quirks to perpetrate harmful conspiracies and create unrest. This can be seen in one instance with the spread of anti-vaccine rhetoric by Russian bots and trolls (Broniatowski et al., 2018), which became especially harmful during the Covid-19 pandemic. Extreme views on topics such as immigration have had a resurgence, in part attributed to the rise of far-right communities (Rauchfleisch & Kaiser, 2020). Overall, a lot of political unrest can be attributed to these pipelines and the echo chambers they create.

I believe the algorithmic radicalization pipelines and its effects can best be understood if viewed through actor-network theory. In order to know what we can do to combat radicalism, we need to understand the relationship between the technology, the communities it touches, the communities it creates, and all the effects of these

communities. If we understand this network more clearly, we will be in a much better position to know what we can do to disrupt it.

Conclusion:

My technical deliverable for this research project will be a google chrome plugin that effectively identifies misinformation using machine learning techniques and displays it to the user in an easily digestible way. My STS research should provide a better understanding on how exactly people become radicalized through their social media usage and what overall effects this widespread radicalization has. I hope I will be able to discover and propose some ways to disrupt the pipeline and decrease the number of people who fall into the algorithmic trap, and that my technical deliverable will provide an easy option for those looking to sift through misinformation.

Refernces:

Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2020). Fake news detection using machine learning ensemble methods. *Complexity*, *2020*, e8885861. https://doi.org/10.1155/2020/8885861

Branson-Potts, H., & Winton, R. (2018, April 26). *How Elliot Rodger went from misfit mass murderer to "saint" for group of misogynists—And suspected Toronto killer*. Los Angeles Times. https://www.latimes.com/local/lanow/la-me-ln-elliot-rodger-incel-20180426-story.html

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., Quinn, S. C., & Dredze, M. (2018). Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *American Journal of Public Health*, *108*(10), 1378–1384. https://doi.org/10.2105/AJPH.2018.304567

Champion, A. R. (2021). Exploring the radicalization pipeline on youtube. *The Journal of Intelligence, Conflict, and Warfare*, *4*(2), 122–126. https://doi.org/10.21810/jicw.v4i2.3754

Faris, R., Hal, R., Bruce, E., Nikki, B., Ethan, Z., & Yochai, B. (2019, December 17). *Partisanship, propaganda, and disinformation: Online media and the 2016 u. S. Presidential election | berkman klein center*. https://cyber.harvard.edu/publications/2017/08/mediacloud

Matthews, J. (n.d.). *Radicalization pipelines: How targeted advertising on social media drives people to extremes*. The Conversation. Retrieved March 24, 2022, from

http://theconversation.com/radicalization-pipelines-how-targeted-advertising-on-social-media-drives-people-to-extremes-173568
Pérez, J. (2019). Handsfree for Web: A Google Chrome extension to browse the web via voice commands. *Proceedings of the 16th International Web for All Conference*, 1–2. https://doi.org/10.1145/3315002.3332443
Rauchfleisch, A., & Kaiser, J. (2020). The German Far-right on YouTube: An Analysis of User Overlap and User Comments. *Journal of Broadcasting & Electronic Media*, *64*(3), 373–396. https://doi.org/10.1080/08838151.2020.1799690
Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2020). Auditing radicalization pathways on YouTube. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 131–141. https://doi.org/10.1145/3351095.3372879
Von Behr, I., Reding, A., Charlie, E., & Gribbon, L. (n.d.). *Radicalization in the digital era*. Retrieved March 24, 2022, from https://www.rand.org/randeurope/research/projects/internet-and-radicalisation.html