Evaluating detection mechanisms for misinformation spread on social media

A Thesis 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

Pablo Weber
Spring 2021

Technical Project Team
Members
None

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____________________________________ Date__05/14/2021___
Pablo Weber

Approved____________________________________ Date______________
Richard Jacques, Department of Engineering and Society

Approved Aaron Bloomfield________________________ Date__11/08/2020___
Aaron Bloomfield, Department of Computer Science

Introduction

General Research Problem

How can we curb the spread of political misinformation on social media?

On December 4 2016, a man named Edgar Maddison Welch arrived at the Comet Ping Pong pizza store in Washington D.C. and fired three rounds from an AR-15 rifle (Ortiz). He explains that he was trying to "save the children", after having read posts on social media stating that the restaurant was harboring sex slaves. He was referring to Pizzagate, a debunked right-wing conspiracy theory claiming that several high-ranking Democratic Party officials and U.S. restaurants were part of an alleged human trafficking and child sex ring. This misinformation - false information deliberately spread (usually on social media) to influence people's thoughts, or more commonly referred to as fake news - led someone to take criminal action in the physical world. Worryingly, in 2014, 61% of millennials in the US claimed to get their political news from Facebook, compared to just 44% claiming to get their political news from CNN (Moon). Clearly, fake news is an incredibly powerful tool for people looking to manipulate and influence people's thoughts.

Technical Project

How can a Chrome extension identify and alert users of posts spreading political fake news on Twitter?

Over the last 16 years, the number of worldwide internet users has increased from 413 million in 2000 to over 3.4 billion in 2016 (Roser, Ritchie, Ortiz-Ospina). This surge in internet usage and accessibility has brought with it a sharp increase in the use of social media websites, such as Twitter and Facebook, the latter now boasting more than 2.7 billion monthly active users, or 34.6% of the world's population as of 2020 (Clement). While social media offers many benefits, such as connecting with distant friends or family members, it has also contributed to a massive rise in disinformation. An example of this is the Russian intervention in the 2016 US presidential election, with “bots” spreading misinformation to millions of Americans in hopes of swaying the election to the candidate that was more accepting of Russian policies and practices.

The political fake news just described is the basis for my technical project which aims to reduce its use. More specifically, my aim is to build a Google Chrome extension that detects Tweets spreading political misinformation. This political fake news detector would rely on cross-referencing Tweets with various articles from different reputable news outlets, like CNBC, CNN, BBC, ABC etc. A machine learning program would then scan through the Tweet to obtain the general gist of the article, request articles from the news outlets with API calls, and cross-reference the Tweet and the articles to get a similarity rating. From this, the program would assign an accuracy score to the Tweet. The ideal end product of this project is a platform that would be available for download on the Chrome webstore, that overlays a small badge at the top of each Tweet with an accuracy score. This would allow users to quickly see if the information being shared by the Tweet is accurate or not.

STS Research

What detection methods can we use when it comes to political Fake News, and how can we mitigate its spread on social media?

In 2016, a few countries played a part in the interference of the US presidential elections. Notably, Russia was found to be one of the biggest actors (Ross, Schwartz, Meek). It is now known that the Russian Internet Research Facility (IRA) based in Saint-Petersburg, Russia, created thousands of social media accounts that purported to be Americans supporting radical right-wing political groups in hopes of promoting the Trump re-election campaign. This “troll farm” is thought to have reached millions of Americans between 2013 and 2017 (Hindman). Clearly, using fake news to advance a political goal can have immense effects. Throughout this paper, I will outline ways in which fake news can be detected and outed, in hopes of putting an end to the spread of disinformation.

First of all, it is worth noting that this phenomenon in not a new one. There have been many cases in the past where fake news has been used to further a political goal. One can think of the "yellow journalism", grandiose coverage of mundane events, that strongly contributed to the start of the Spanish-American War, or the fake stories spread by the Associated Press (AP) that ultimately contributed to the presidency of Rutherford Hayes, or more recently the fake news stories spread by the Soviet Union to the American people during the Second World War (Hindman). In the past, fake news was a great way for the press to further a goal. If they could convince their readers of an idea, they could potentially sway their political opinion. However, nowadays, this technique is not only limited to be used by the press.

With the rise of social media websites where anyone can post whatever they would like, people can now reach potentially millions of other people. This is unprecedented. While this comes with many benefits, it also comes with many pitfalls, most notably the now *widespread* use of fake news. This would not be as concerning as it is if fake news was not an effective political tool, however research suggests that it is. Because social media is now the most used medium for getting news, many people unknowingly find themselves in filter bubbles. This is where an individual is only shown news articles that "interest" him (according to social media content algorithms), which usually means they are only exposed to news articles that align with their political views. Because of the lack of political diversity and adverse opinion, the individual only strengthens their political belief. This also leads to the radicalization of individuals as their filter bubble gets smaller and smaller. Research has shown that this self-picked news influences public opinion more than traditional, balanced exposure to news (Garrett, Stroud).

The use of social media as fake news outlets is also powerful because of the phenomenon known as social proof. This is when individuals are more likely to believe an idea or concept when many other people do too. It is inherently human to want to believe what others believe, because we assume that they already did the guesswork, and also because it is just cognitively easier to lean on the beliefs of others. This is even more pronounced when the "others" are people we know and trust, like family members or close friends. One last reason why the spread of fake news on social media is so powerful is because of the frequency of posts. On social media, the same news article might be shared by millions of people, and could thus show up on your timeline multiple times. This repetition alone can make the news stories more believable.

While real people do indeed spread fake news on social media, when an actor wants to effectively weaponize fake news, they more often than not use bot accounts. These are automated accounts that are usually made to look like regular, human-run accounts. The reason bot accounts are used is because they allow for disinformation to be posted by potentially millions of accounts, all automated, which allows for a much wider reach compared to a single human.

"Troll farms", networks of bot accounts, is the primary method actors use to spread fake news on Twitter. Thousands of fake accounts are created, purporting to be real human beings, that spread some type of information to further a goal, usually political. So far, no social media websites have implemented a catch-all tool to identify political fake news. While these accounts seem human-like on the superficial level, if we analyze the details of their posts, we notice that they tend to demonstrate peculiar behavior, which distinguishes themselves in a few key ways:

First, these fake accounts are almost always grouped into clusters, where within the clusters they Tweet similar information and follow overlapping accounts. What's more, the accounts have a tendency to Tweet at the same time, a phenomenon commonly referred to as a spike. These spikes usually take place at regular intervals throughout the day. While the typical Twitter user is active around the time before work, during their lunch break, and after work, these fake accounts tend to tweet at exact times – ie. 8:50am before work, or 5:20pm after work.

Secondly, contrary to popular belief, these fake accounts are not linking to small, obscure websites. In fact, it was found that most of these accounts are actually linking to a few large, "reputable" conspiracy websites, with 79.3% of tweets from these fake accounts linking to just 24 news outlets (Hindman). This amounts to over 6 million Tweets linking to only 24 websites in the month before the 2016 election (Hindman).

Finally, these fake accounts tend to follow other fake accounts that are grouped with them in a cluster. Not only that, they tend to reference the other accounts, often re-tweeting and interacting with their Tweets. In fact, a study found a positive correlation between the number of users a fake account followed within their cluster and the amount of disinformation spread on Twitter in the month before the 2016 election (Hindman).

This un-human-like behavior can thus be used to spot accounts that comprise Troll Farms. First, one can refer to the content and the frequency of posts to identify clusters. Analyzing the actual content of the Tweet allows us to first identify accounts that are currently spreading fake news.

The next step is to identify fake accounts. As previously explained, these accounts tend to post at the exact same time every day. Analyzing the post time of accounts spreading misinformation allows us to identify suspected fake accounts. Once we have identified accounts posting at the same times every day, we can look to the accounts they follow. This will show that there is typically a strong intra-following between cluster members. This will not only further identify suspected fake accounts, but also bring other fake accounts into the spotlight. From there, one can look at the time of posting, as well as the interval between posts.

Finally, the outlets linked in the Tweets can be analyzed to determine whether the account is constantly linking to the same, controversial news outlets. If these are all determined to be true, there is a, not certain, but high likelihood that the account is a fake account used to spread disinformation.

These are the primary ways in which we can identify the bot accounts, however there are other attributes that these accounts portray such as posting the same article multiples times, no biographical information (no bio), very high number of tweets, same follower to following number, and not actually posting any original content, only retweeting articles. Clearly, fake news is a problem, especially in today's political climate. With actors such as Russia and Iran with virtually unlimited resources, the spread of misinformation meant to sway the elections is unquestionable. This has led to unpredictable consequences as seen during the 2016 elections.

Consequently, identifying fake news, whether it be individual posts or the source of the problem, the fake accounts spreading it, will help users of social media have a more neutral point of view on political matters. Furthermore, it will encourage users to do their own research, form their own opinions, and will contribute to a dramatic decrease in the amount of radical misinformation being distributed. Equally important, it will decrease the exposure to conspiracy websites where people are easily influenced that will support a more stable political climate.

References

Roser, M., Ritchie, H., & Ortiz-Ospina, E. (2015, July 14). Internet. Retrieved from https://ourworldindata.org/internet

Clement, J. (2020, August 10). Facebook: Active users worldwide. Retrieved from https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Moon, A. (2017, September 08). Two-thirds of American adults get news from social media: Survey. Retrieved from https://www.reuters.com/article/us-usa-internet-socialmedia/two-thirds-of-american-adults-get-news-from-social-media-survey-idUSKCN1BJ2A8

Ortiz, E. (2017, June 22). 'Pizzagate' Gunman Edgar Maddison Welch Sentenced to Four Years in Prison. Retrieved from https://www.nbcnews.com/news/us-news/pizzagate-gunman-edgar-maddison-welch-sentenced-four-years-prison-n775621

Hindman, M., &amp; Barash, V. (n.d.). Disinformation, 'fake news' and influence Campaigns on Twitter. Retrieved from https://knightfoundation.org/features/misinfo/

Ross, B., Schwartz, R., Meek, J. (December 15, 2016). Officials: Master Spy Vladimir Putin Now Directly Linked to US Hacking. Retrieved from https://abcnews.go.com/International/officials-master-spy-vladimir-putin-now-directly-linked/story?id=44210901