Identifying political misinformation on Twitter with a Chrome extension

(Technical Paper)

Evaluating detection mechanisms for misinformation spread on social media

(STS Paper)

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
Fall 2020

Technical Project Team Members
CS Technical Project will be completed in the Spring

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

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.

Identifying political misinformation on Twitter with a Chrome extension

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, colloquially referred to as Fake News - especially prevalent during the US presidential elections. 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.

Recently, Twitter has implemented a feature aimed at combating a portion of fake news concerning the Coronavirus. It scans through any tweets that are thought to be related to the Coronavirus, and overlays a warning message if the Tweet is thought to conflict with the CDC guidelines regarding public safety. This is similar to what the Chrome extension would do. However, detecting fake news regarding the Coronavirus is more trivial, as there is a centralized authority – the CDC – that Twitter can reference. However, when it comes to fake news in the political space, there is no official “authoritative political entity” like the CDC.

For this reason, the political fake news detector would need to rely on cross-referencing Tweets with various articles from different reputable news outlets, like CNBC, CNN, BBC, ABC etc. A machine learning program would 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. Indeed, this can’t just be an on/off switch, but rather a percentage, simply because some sections of the Tweet could be accurate, and some not so. The program would also check the reputability of the news outlet the Tweet is referencing, and take that into account. Obviously, outlets that are known to be controversial and spread fake news, like Alex Jones’ InfoWars, will more often than not have a lower accuracy score. Finally, users would be able to enter their advice concerning whether the Tweet is factually accurate or not. This would be represented as a “thumbs up” or “thumbs down” button,

where the program would then take into account the number of votes for both sides of the equation. Of course, this would only constitute a small portion of the rating because this could easily be abused, but it is useful feedback nonetheless.

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. The badge would be in the top right of each Tweet, and its color would depend on the accuracy score. Tweets that have been judged to be in the threshold of not accurate after having been cross-referenced to reputable news articles and user feedback, or in other words inaccurate Tweets, would have a red badge, while moderately accurate Tweets would have an orange badge, and accurate Tweets a green badge. This would allow users to quickly see if the information being shared by the Tweet is accurate or not. This accuracy score would be automatically calculated using machine learning whenever a user scrolls past a Tweet, or clicks on a Tweet.

Evaluating detection mechanisms for misinformation spread on social media

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 - or 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 campaign.

This “troll farm” is thought to have reached millions of Americans between 2013 and 2017 (Hindman).

This is the primary method actors use to spread fake news. 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.

These fake accounts are grouped into clusters, and 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, called a spike. These spikes are usually taking place at regular intervals. This is obviously not similar to human behavior. 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. Because these clusters send out Tweets at the same time, it makes it easy to spot.

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). Thus, monitoring which users are spreading and linking to these websites offers a way to catch and penalize fake accounts.

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 amount 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).

Consequently, one can use the content and the frequency of posts to identify clusters. These links are further confirmed when there is strong intra-following between the cluster members. 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. Identifying these accounts is a key part in stopping the spread of political fake news on social media.

Conclusion

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. Therefore, 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. 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. What's more, it will lead to a decrease in the exposure to conspiracy websites, where those people easily influenced are swayed, and will result in a more stable political climate.

Word count: 1757

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., & 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