Method for Preventing the Spread of Damaging False Information on Social Media
(Technical Paper)
Comparison of Solutions and Understanding Effects & User Behaviors
(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

Eric Stoloff
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 Eric Stoloff . Date 10/24/2020 .
Author Name
Approved _____. Date _____.
Technical Advisor: Raymond Pettit, Assistant Professor, Department of Computer Science

General Research Problem

How does damaging false information spread through social media and is there a way to prevent this that strikes a balance between being effective, practical, and ethically sound?

The spread of false information has become a major problem in our society today. Although mainstream news networks are not without incident, social media is the most important medium through which damaging false information propagates. Social media is defined as "any online or digital medium provided and/or collected through a channel that enables the two-way sharing of information, involving multiple parties. This includes social networking sites, texting, blogs, etc." (DHS, 2018). In 2016, social media was the most important information source for 51% of people (Newman et al., 2016), and its popularity and use continue to grow every year (Chaffey, 2020). This growing user base presents a growing opportunity for the spread of false information because social media allows people to not just receive information from news and other networks, but the ability to create their own media and the means to potentially spread it worldwide (Dafonte-Gomez, 2018). Although social media and the internet are relatively young technologies, there are already a multitude of examples that illustrate the consequences of the proliferation of false information on social media, including the 2016 American Presidential Election (Allcott & Glentzkow, 2017). In the three months before the election, there were over 156 news stories that circulated on social media that were determined to be false, the vast majority of which favored Donald Trump (Allcott & Glentzkow, 2017). Not only were there more instances of false information biased towards Trump, but these stories were shared over four times more than false information supporting Hillary Clinton. Although it is possible for an individual or group to create false information that serves a purpose, the case of the 2016 election can be explained with the advertisement revenue that the creators of many of the false stories received (Allcott & Glentzkow, 2017). After initial circulation, these stories were

correctly identified as fake and flagged, but the major problem in preventing the spread of false information is its scale. There are now 3.96 billion social media users and 4.57 billion total users of the internet worldwide (Chaffey, 2020). The issue with finding cascades of false information is that all the content generated from these users is an impossibly large pool of data that would take incredible amounts of computing power to parse through. In addition, the vast majority (about 90%) of information available is credible information, so it is easy for false information to naturally hide (Kumar & Shah, 2018). Solutions that will expose false information will need to be effective while also being practical enough to be implemented within real computer networks. The final factor in determining how favorable a possible solution will be is the ethical implications surrounding what defines false information and the level of intrusion into personal information that may occur in the process of finding false information. Given that people have personal liberties like Freedom of Speech in the U.S., there is some question as to whether the Government or social media networks have the right to prevent people from sharing through social media (Maitland & Lynch, 2020). However, false information, especially that which is sneakily presented as fact, is too damaging to let go unchecked.

Method for Preventing the Spread of Damaging False Information on Social Media

What is a concrete method that prevents social media users from sharing false information which creates a belief in something that is inherently flawed and addresses the issues with current solutions?

The idea of misinformation or false information is not new to society. It has been used for thousands of years as a means to confuse, manipulate, or otherwise grow a belief in a falsity. However, social media is a young technology as it relies on the internet. It is still changing social

norms and scientists still do not fully understand the impact that it will have (Dafonte-Gomez, 2018). The factor that has changed with social media is the degree to which false information can spread and influence people's decisions. False information can be dealt with in isolated instances, but it can become dangerous when millions of people believe in something erroneous. For example, a fake video titled "Somalis Pushed Into Shallow Grave Ethiopia" resulted in violent conflict between two ethnic groups in Ethiopia, perfectly illustrating the severity of the problem (Guo et al., 2020). People believed the information with so much conviction that they were willing to commit acts of violence over it. My project will address this by developing a way to stop or slow the spread of false information on social media that can be realistically implemented by social media sites.

The main problem with social media that makes it harder to eliminate false information is the sheer diversity of content that social media contains. While news networks provide reports of real-world events and other information that is expected to be factually correct, social media gives people the freedom to post almost anything they want, including their opinions, jokes, and feelings alongside actual information. This makes it much harder to discern what category a post falls into. The most common way in which people do this is by that format that the information is presented in (Kumar & Shah, 2018). If something looks informal or hand-written, it is not very credible to users. However, if a post is presented with a genuine, formal look, its credibility increases dramatically (Kumar & Shah, 2018). Due to this and factors like the recency of the information, humans are terrible at distinguishing between true information and false information (Westerman & Spence, 2014). This is supported by many studies, including one where people correctly identified the hoax in a set of articles only 66% of the time (Kumar & Shah, 2018).

With social media users having this low of an accuracy level picking out false information, it is no surprise that it is able to proliferate so easily.

This problem has not gone unnoticed, as there are many people around the world that are currently attempting to solve it. The solutions that currently exist can be divided into four categories, the first of which is content-based detection, or classifying information based on visual and textual data scraped from social media (Guo et al., 2020). The second category is social-context-based detection, which deals with how users interact with each other to share information. The third category is future-fusion detection, which uses elements of both previous categories. The final category is deep-learning detection, which works by using neural networks to learn the information. Although these solutions are somewhat effective, they have two main drawbacks. The first problem is that they cannot find false information in a timely manner, which is extremely important given how fast social media posts can go viral (Guo et al., 2020). In addition, they are binary in that they determine if some information is true or false. They do not give any reasoning for their decision, which is needed to un-sow the seeds of belief that the false information may have planted (Guo et al., 2020; Kumar & Shah, 2018). These drawbacks seem to suggest that external algorithmic-based detection may not be the best route in the fight against the spread of false information.

This can be addressed by developing a method that can be "built-in" to social media applications, and their user interfaces. The project will demonstrate the solution by communicating with the well-documented application public interfaces (APIs) of top social media sites like Facebook, Twitter, and YouTube. In addition, there are many publicly available datasets of social media data available on websites like Kaggle. It will not rely on a complicated algorithm, but will ideally prevent the spread of false information by helping users to recognize it

with a greater accuracy than humans are naturally capable of. Once the solution is developed, it will be tested on a variety of users. After this testing is complete, there may need to be changes to the solution, but hopefully the finished product will be successful in stopping the spread of false information on social media while being a future model for solutions that work by improving the user's ability to distinguish between the two types of information.

Comparison of Solutions and Understanding Effects

What are the factors at work in the spread of false information on social media and how can we converge on a solution that is effective in preventing its spread, practical in implementation, and ethical towards users' privacy?

Introduction

The expansion of false information through social media is a major problem in modern times with social media continuing to expand annually. As users' capacity to participate in the dissemination of information has increased due to social media, users have stepped up to fill this capacity (Dafonte-Gomez, 2018; Westerman & Spence, 2014). In doing so, users may take part in the spread of misinformation. To stop this, it is important to evaluate current solutions with appropriate criteria. In doing so, it is necessary to study the relationship between the user and information, and the relationships between groups of users (Bin et al., 2020). The current gap in the understanding of the motivations behind users sharing information is that many studies assume that the users have read and understand the information they share, yet this is frequently not the case (Maitland & Lynch, 2020). The human brain produces reward stimuli from the simple act of sharing regardless of the content, so it is necessary to understand how this type of false information sharing can be prevented (Dafonte-Gomez, 2018). Another area that will be explored is the advances in machine learning and natural language processing that have occurred

recently. Although there are limitations to algorithmic approaches to detecting false information, these advances may increase the effectiveness of these solutions and make them viable choices in the future (Bin et al., 2020; Kumar & Shah, 2018). A third area that must be accounted for is the ethical implications that solutions to this problem may have with respect to compromising personal information and violating the rights that individuals possess (Maitland & Lynch, 2020). Overall, this research project will delve into the root causes that stimulate the spreading of false information and it will create criteria that potential solutions can be evaluated on in order to compare them.

Actors in the System of Information Spreading on Social Media

By its very nature, social media is a complex, interconnected web of information and connections between users. Although users, people, and society are obvious ones, there are many more actors that play a role in the context of social media. The first of these is government agencies, which interact with social media in passing important information to the public. For example, FEMA uses social media to keep the public informed with emergency information during disasters (Lindsay, 2011). Agencies also interact with social media by getting information from it and watching over various platforms to find illegal activities. This is because the government has recognized that harnessing the power of users can benefit them and they can gain "situational awareness" from information cascades on social media (Lindsay, 2011). Another major entity in the context of social media is the media companies themselves. Because they directly run and control the social media applications themselves, they have a huge role in mitigating the spread of false information and protecting user privacy. A third group of actors to consider with regards to social media is the mainstream media and news. They are important because they monitor behavioral trends of users and are an integral part of fact-checking

information that passes through social media (Dafonte-Gomez, 2018). The final entities to consider in this context are the bad actors, or those who would create false information and spread it. These could include individuals who act alone, or organizations with a primary goal, and their capabilities could vary from a single account to a whole botnet of accounts under their control (Kumar & Shah, 2018). For example, the teenagers living in the town of Veles, Macedonia who ran sites that posted most of the "fake news" relating to the 2016 American Presidential Election (Allcott & Glentzkow, 2017). Although there are some other related actors that play a role in social media in general, like marketers advertising to users, they are not relevant to the spread of false information.

Addressing the problem of false information propagating through social media is a major challenge for our society today, and it is one that will continue to present itself in the future. Although solutions have been developed in the areas of propagation modeling, detection from user data/metadata, detection using text, image, and video content, and machine learning, the only result that is known is how effective they are (Bin et al., 2020; Kumar & Shah, 2018). Specifically, the solutions are measured on the percentage of false information that they find. One issue that stems from this is the lack of consistency in datasets used to test these solutions. Although the effectiveness of the solution is important, the wide variety of social media and content used makes it hard to compare solutions to each other (Bin et al., 2020; Kumar & Shah, 2018). In addition, the solutions are tested with relatively smaller groups of data compared with the entire breadth of social media. The first problem posed by this is that the expertise of the solutions is very narrow. For example, if a machine learning algorithm was trained to recognize false information on a dataset of political posts, it may not be the most proficient at recognizing false information relating to the species of marine life inhabiting the Great Barrier Reef. This

problem is multiplied by the incredible diversity of content on social media (Bin et al., 2020). Secondly, there could be complicated solutions that work for small sets of data but have a bad asymptotic running time. This means that as the size of the input increases, the running time the algorithm takes to complete approaches infinity, so these solutions could not be scaled for use with the ocean of data that is social media. Because of this, scalability and practicality of implementation are necessary to evaluate methods of preventing the spread of false information. The final important aspect of solutions that is not taken into account when existing solutions are compared is ethical and privacy concerns. In the process of finding false information, it matters how much user personal information is exposed because this creates an imbalance of power towards those possessing the information that has the potential to be used maliciously (Maitland & Lynch, 2020). Adding to this, censorship is a slippery slope, so solutions must focus on identifying false information while not infringing on rights like Freedom of Speech/Expression. The addition of these criteria to comparing solutions to this problem will allow for a much better understanding of their suitability for actual implementation by social and news networks.

Besides the case of botnets, or many non-human accounts controlled by a group or single person, false information spreading through social media is done as a consequence of user behavior. The solutions to this problem will use technology to counteract this unwanted behavior, so a social construction of technology (SCOT) model will result.

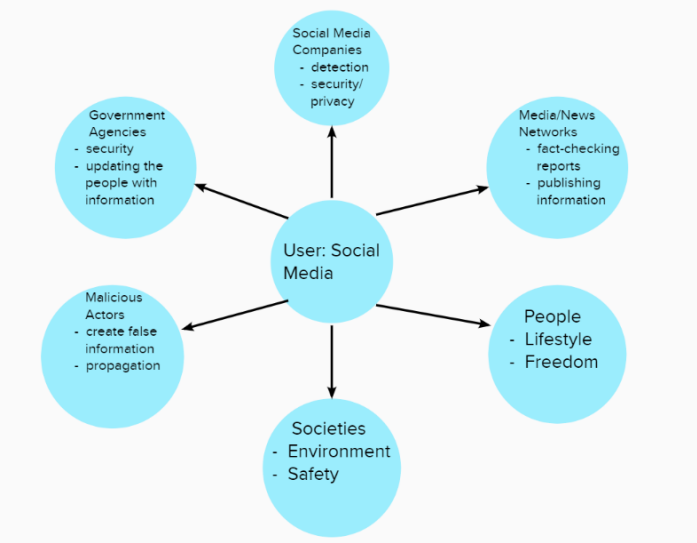

Figure 1: False information detection in social media SCOT model: The user sits between all other entities and is the medium through which information passes.

This model can be used to show that the success of a method for preventing the spread of false information is defined by criteria that is defined by the groups and stakeholders that play a role in it. The model is set up with the user in the center because in terms of the spread of false information and detecting it, every exchange of information goes through users. False information is created by malicious actors, and it spreads through users sharing content, even if the actors use things like botnets to help out. Then, the false information can do what it was intended to, influencing people's beliefs and the environment of our society (Heath et al., 2014). The detection of false information also occurs through users, as user content, metadata, and relationships to other users is utilized by various means of detection by Social Media Companies and Media/News Networks (Bin et al., 2020). Government Agencies exchange information with the user in two ways, with agencies simultaneously being informed by users and providing important information to the public (DHS, 2018). The user being at the center of the SCOT model suggests that while solutions will involve all entities, users ultimately have an important responsibility in preventing the spread of false information (Westerman & Spence, 2014).

Evidence/Data Collection

My STS Research project will be a scholarly article that organizes the literature on how false information spreads through social media, the effects it has, and evaluates current solutions to this problem on diverse criteria. As previously stated, these include effectiveness, practicality and scalability, the degree to which the solution threatens user privacy, and speed in which the solution is able to work. The sources for this research will be the original sources describing solutions that impede the spread of false information and they will be evaluated on all these criteria and compared. The issue of the datasets not being standard can be worked around by paying close attention to the unique characteristics of each dataset in order to infer how they may have affected the result of testing these solutions.

Methods

In an effort to understand as much as possible about the effects that the proliferation of false information on social media has, as well as how it occurs, I will explore sources that cover the topic with respect to different areas. For example, I will delve into the sociology and psychology behind users' interaction with social media so that the paper can provide an adequate explanation of the motivations users have when sharing information and the effect that false information has on society. The basis for analysis of potential solutions for this research will be the degree to which the solutions satisfy the criteria. This research will be explanatory case research, where cases will be chosen to be representative of the range of possible solutions to this problem.

After potential solutions are evaluated according to the list of criteria, it will hopefully be possible to identify the best of the solutions. If one solution does not emerge as the clear choice, then this research will still be sufficient to determine the benefits and drawbacks of each one

examined. Whichever the case may be, this research will be useful to social media companies and media/news networks in helping them decide which direction to take in the fight against the propagation of false information on social media.

Conclusion

Social media is one of the most important ways to interact with information in our modern society. Its user base will continue to follow growing trends, and social media will become even more ingrained in the flow of information in the future. With that said, it is all the more important to ensure that false information cannot proliferate to the degree that it does on social networks today. Although this project lacks the budget and technological capabilities to be a perfect solution, my tightly-bound research and technical project will ideally advance the prevention of false information spreading through social media in a positive direction. The technical project may not result in 100% accuracy, but it will be successful in mitigating some spread of misinformation. The STS research has great potential for future work because solutions will be evaluated on benchmarks that, so far in my research on this topic, have not been evaluated on. This will provide a means in which different types of solutions can be compared and a way in which their fitness for a more specific false information detection problem can be determined.

References

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives, 31(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, & Zhiwen Yu (2020, August 1). The Future of False Information Detection on Social Media: New Perspectives and Trends. ACM Computing Surveys, 53(4), 68 - 103.

Chaffey, D. (2020, August). Global Social Media Research. Smart Insights. Retrieved from https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/

Dafonte-Gómez, A. (2018). Audience as Medium: Motivations and Emotions in News Sharing. International Journal of Communication (19328036), 12, 2133–2152.

Department of Homeland Security. (2018). Countering False Information on Social Media in Disasters and Emergencies. https://www.dhs.gov/sites/default/files/publications/SMWG_Countering-False-Info-Social-Media-Disasters-Emergencies_Mar2018-508.pdf

Heath, D., Singh, R., & Ganesh, J. (2014). Social Media at SocioSystems Inc.: A Socio-technical Systems Analysis of Strategic Action. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6758675

Kumar, S. & Shah, N. (2018). False Information on Web and Social Media: A Survey. Social Media Analytics: Advances and Applications, by CRC press, 2018. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Lindsay, Bruce R. (2011). Social Media and Disasters: Current Uses, Future Options, and Policy Considerations. Federation of American Scientists. https://fas.org/sgp/crs/homesec/R41987.pdf

Maitland, N. & Lynch, J. (2020, March 1). SOCIAL MEDIA, ETHICS, AND THE PRIVACY PARADOX. Journal of Internet Law, 23(9), 3 - 14.

Newman, N., Fletcher, R., Levy, D., & Nielsen, R.K. (2016). Reuters Institute Digital News Report 2016, Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/research/files/Digital%2520News%2520Report%25202016.pdf

Suliman Aladhadh, Xiuzhen Zhang, & Mark Sanderson (2019, February 11). Location impact on source and linguistic features for information credibility of social media. Online Information Review, 43(1), 89 - 112.

Westerman, D., Spence, P.R. and Van Der Heide, B. (2014), Social Media as Information Source: Recency of Updates and Credibility of Information. J Comput-Mediat Comm, 19: 171-183. doi:10.1111/jcc4.12041