A SYSTEMATIC APPROACH TO ANALYZING SYNDICATION OF VIRAL NEWS STORIES

A PREPRINT

Gaurav Deshpande
Institute for Software Research
Carnegie Mellon University
Pittsburgh, PA 15213
gdeshpan@andrew.cmu.edu

Sam E. Teplov
Institute for Software Research
Carnegie Mellon University
Pittsburgh, PA 15213
steplov@andrew.cmu.edu

Alon Peer
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
apeer@andrew.cmu.edu

Dr. Nicolas Christin
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
nicolasc@andrew.cmu.edu

December 5, 2019

ABSTRACT

In this work, we present a novel, semiautonomous methodology for analyzing syndicated news articles. We first detail our manual process of news syndication analysis from which we observe a number of unsettling patterns. From there, we go on to categorize and define two different forms of syndication: top-down and bottom-up. We then transform our manual process into a semiautomated one which combines web scraping and machine learning techniques to create and then analyze a dataset of viral news stories. From the analysis, we are able to make some inferences about the nature of syndicated news articles, as well as trace back a viral news story to its original publisher and article. By presenting this work, we hope to minimize the effects that blind news syndication has on the spread of misinformation.

***K*eywords** Fake news · News propagation · News syndication

1 Introduction

News are everywhere- television, newspapers, social media. We consume it whether we like it or not. But the question is, how do we know which news are true, and which are false?

Fake news has been a real issue since the spark of social media, especially in the recent years, with its spike coming during the 2016 U.S. presidential election. In fact, a study from Ohio State University shows that fake news had a significant impact on the results of the elections. The researchers found a correlation between the amount of Obama voters who believed fake news about Clinton and the amount of them who chose to vote for Trump. At that time, the media was swarmed with fake news, and one could hardly blame people who couldn't distinguish them from real news.

But what exactly is fake news? According to Wikipedia, "fake news is a form of news consisting of deliberate disinformation or hoaxes spread via traditional news media (print and broadcast) or online social media". In addition to this definition, fake news may also include clickbait stories, propaganda, satire, biased news, or even sometimes just bad journalism.

These days, anyone can publish content on the internet, whether it is on a website, a blog, or a social media profile, and potentially reach large audiences. The fact that people nowadays rely on the internet to consume news doesn't make this problem any easier to solve. Fake news is very profitable for the publisher, as viral stories generate large

amounts of traffic, thus increasing advertisement revenue. For that reason, fake news are widely common around the internet. Another motive to publish fake news is to promote a political agenda. Fake political news stories are meant to persuade consumers to accept a biased of false beliefs (like the fake news regarding Hillary Clinton during the 2016 U.S. presidential election). This is the main reason detecting fake news is such a challenge. These false stories were created to sound true and interesting.

A different variation of fake news is syndication. A website can take a real news story, and change it a bit, so it will fit that website's beliefs. A news story can be syndicated in various ways- changing the title, adding interpretations that aren't necessarily true, and much more. A popular form of syndication is news stories that are posted by local newspapers and websites, and as they get viral, these stories get picked up by bigger and more popular news websites. On the way from the local news source to the national (or international) news website, the story can get syndicated. Sometimes, it's not necessarily on purpose, but it happens due to the lack of due diligence by the mainstream media.

One can try and detect fake news by checking the source of the story, read beyond the title (which is intended to attract users and thus contains more interesting and controversial topics), try to trace the article back to the original source, or even make sure the article is not a satire by any chance. Even though it's possible, doing so is not an easy task, and that is why there has been a lot of work put into automating this process. A step towards this, is detecting and analyzing the syndication of viral news stories. That is, how stories change through time and the difference between articles regrading the same stories but posted in different websites.

2 Background

The field of misinformation and fake news has become widely researched by academics ever since the Russian misinformation campaign against the 2016 United States presidential election came to light[1]. The most straightforward approach to detecting fake news is applying various machine learning techniques, and this is what many researchers have focused on thus far.

2.1 Fake News Detection

Shu et al. goes through a comprehensive overview of all of the various ways of attacking the problem of fake news detection and classification from a data mining perspective [2]. From the perspective of detecting fake news via machine learning techniques, Shu et al. outlines that there are generally four different types of features that can be used to detect fake news: news content features, linguistic-based features, visual-based features, and social context features. These features are then applied to one of two different types of models: news content models, or social context models . Just as they sound, news content models focus on fact checking claims in the news article or looking at stylistic features that may indicate that the article is fake. Social context models, on the other hand, focus on how users interact with the article, the virality of the article, and the propagation path of the news story. These two different methods are rarely used independently, but rather compliment each other in many hybrid approaches, such as ensemble models [2].

2.2 Open Issues in Fake News Detection

Shu et al. closes by giving an overview of many of the various open areas of research in this field, as well as proposing some future research ideas. Shu et al. outlines four future research directions: data-oriented, feature-oriented, model-oriented, and application-oriented. Within the area of data-oriented research, Shu et al. brings forward a couple of issues that need to be addressed. For one, there is no large, encompassing fake news data set which can be used as a benchmark to facilitate future research [2]. In fact, this was such a large problem that in 2018, Shu et al. published another work detailing their methodology for building a large scale, open source, data set consisting of fake news, which is now available on Github [3]. Another idea that Shu et al. proposes for future research within the data-oriented direction is early fake news detection [2]. Many of the current state-of-the-art fake news detection methods rely on data that needs to be accumulated over time and information that is not readily available when news is just starting to go viral [4]. Shu et al. suggests creating a system that provides fake news alerts during the dissemination process so that misinformation can be flagged before it is viewed by a large audience. Another suggestion that is made is to look at social media posts made within some time delay of the original posts as a means of verifying the veracity of the post [2].

3 Related Work

There has been work done on early fake news detection by analyzing propagation paths and time series. Most of this work, however, has been limited in scope in terms of the domains that it covers, as well as the type of media being analyzed. For example, past research has focused on a either a few social media platforms, or a specific media outlet

[4],[5]. The research that has spanned across multiple different domains of news (social media, various news outlets, etc.) has been limited in scope when considering the type of media that is being analyzed [6].

3.1 Domain-Specific Approaches

3.2 Media-Specific Approaches

4 Methodology

4.1 Phase 1 - Manual Inspection

To begin the process of analyzing syndication, we started off by manually gathering articles and seeing the propagation path of those articles. We tried to look for a pattern in the way articles change throughout time. This phase included x steps.

4.1.1 Article Search

To begin the manual inspection. We started looking through the internet for articles that gained a lot of traffic relative to other articles in the same website. We looked for a high number of shares or comments. Some of the news sites we looked at are "Daily Mail", "The Washington Post" and "Daily Caller".

4.1.2 Original Source Search

After we filtered out some articles due to content and popularity, we tried tracing the articles back to their original source. We google key words from the article, and that way we essentially got a list of different websites' articles of that news story. We paid close attention to the dates, to try and find the earliest article regrading said news story.

4.1.3 Propagation path

Once we found the original news source, we looked at how the article title and content change over time and throughout different news sources. We tried to see how the website's beliefs changed the way the story was described in each and every news site, that is, how news got syndicated.

5 Analysis

6 Contributions

7 Findings

8 Future Work

9 Conclusion

References

[1] Constanze Stelzenmüller. The impact of russian interference on germany's 2017 elections. *Testimony before the US Senate Select Committee on Intelligence June*, 28, 2017.

[2] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[3] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 2018.

[4] Yang Liu and Yi-Fang Brook Wu. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[5] Zhiwei Jin, Juan Cao, Yu-Gang Jiang, and Yongdong Zhang. News credibility evaluation on microblog with a hierarchical propagation model. In *2014 IEEE International Conference on Data Mining*, pages 230–239. IEEE, 2014.

[6] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, pages 188–202. ACM, 2018.