Elizabeth Williams

Professor Stacey Suver

CSE 300

30 March 2018

Methods of Detecting Fake News

Fake news consists of fabricated or falsified information that mimics the appearance of real news content. While yellow journalism is nothing new, the advent of the internet and social networking websites where anyone can share content has created a perfect storm for the dissemination of fake news. Fake news can misinform and sway the public on important issues, such as in science and politics. A massive amount of content is posted and shared daily on social networking websites, making it too costly and impractical for humans to assess the credibility of shared content. To tackle the problem of assessing the credibility of large volumes of content and preventing misinformation from going viral, automated methods for detecting fake news must be investigated and implemented. Research on automated methods of detecting fake news thus far has worked on classifying news as fake based on headline contents, user interactions, and knowledge of the crowd.

In “Misleading Online Content: Recognizing Clickbait as “False News”” by Chen et al., the authors examine methods for the automatic detection of deceptive clickbait. Clickbait's purpose is to attract attention and encourage visits to a webpage, and such headlines grab attention and promote sharing, especially on social media, but can be misleading and promote the spread of rumors and fake news. The authors propose a hybrid approach to automating clickbait detection by using text and non-text cues. Text cues are measured through analyzing text at a

lexical and semantical level, as well as at a syntactic and pragmatic level. Lexical and semantic analysis is done by training a classifier to identify the probability of individual words belonging to a prototypical clickbait headline. A potential clickbait headline can then be evaluated by taking the probability of each word in the headline, then evaluating the individual probabilities through Bayes' formula or vector distances. At the syntactic level, the authors claim that forward referencing, combined with use of numbers, disparate topics, and provocative adjectives can bait readers by eliciting curiosity. The authors also investigate and consider the use of non-text cues such as image analysis and user behavior analysis to assess the presence of clickbait. Images incongruent with the headline can indicate misinformation in headlines. The authors thus claim it is valuable to look at how users respond to content, and that higher user interest suggests a lower probability that the content is clickbait. This research proposes multiple, compatible methods that can be used to automatically identify clickbait, which is valuable since massive amounts of clickbait are spread throughout the internet and human verification of clickbait would be time consuming and costly. The research is limited in that it does not offer a concrete implementation yet, nor does it evaluate the effectiveness of these individual and combined methods classify clickbait.

Chen et al. proposed potential methods for automatically identifying clickbait, a common form of fake news, but neglect studying the social networking websites in which fake news commonly appears. In "Automatically Identifying Fake News in Popular Twitter Threads" by Buntain and Goldbeck, the authors develop methods for automating fake news detection on Twitter through classification models. The authors trained models using Twitter credibility datasets, CREDBANK and PHEME. CREDBANK is a large-scale dataset of conversations with crowdsourced accuracy assessments, while PHEME is a curated set of conversations about

rumors with accuracy assessments by journalists. Once trained, they evaluated the models against BuzzFeed's Fake News dataset. The authors found that model based on the crowdsourced dataset CREDBANK outperformed the model based on the journalist sourced dataset PHEME and concluded that CREDBANK models may be more appropriate for automated fake news detection on Twitter. The research demonstrates an implementation of an automated method for the detection of fake news on social media, and also suggests that crowdsourced data may be more effective than data from experts for the task of identifying fake news as it appears on social media. A limitation of the research is the underlying structural differences in the PHEME and CREDBANK datasets, which may have resulted for the difference in feature sets and accuracy in identifying fake news. Additionally, the BuzzFeed Fake News dataset is limited in that it is only 35 political stories. Further evaluation of these models should be done when a more large-scale dataset of fake news becomes available.

Buntain and Goldbeck's work proposed an effective model for automatically detecting fake news without the need for human analysis, but they do not implement their findings to assist general users in navigating fake news content on social media. Their models also do not work on fake news photos or videos. In "Fake Tweet Buster: A Webtool to Identify Users Promoting Fake News on Twitter" by Saez-Trumper, the author proposes a hybrid technique for identifying fake photos of news on Twitter and implements it as a Webtool. The motivation behind this research is that it can be difficult to ascertain the context and validity of photos, especially as social media such as Twitter is used to follow ongoing conflicts. On occasion, such photos have gone viral and have been republished by official news sources. The author developed "Fake Tweet Buster", which guesses the credibility of a Tweet using several criteria, then asks the user to confirm whether it is Fake or Legitimate. The researcher's approach to guessing whether the Tweet is fake

or legitimate uses a reverse image search, user analysis of the Tweet's author, and crowd sourcing. The reverse image search looks for an identical or very similar image using Google Images and TinEye. The reverse image search returns context and temporal information about the image, so this information can be used to guess about if a photo shared in a Tweet is accurately represented by its description. By applying these techniques, the tool can determine if an image said to be of Country X actually refers to country Y, and if the year is accurate. User Analysis is performed by measuring the number of tweets, followers, and age of an account, since new accounts with fewer tweets tend to be more suspicious. Finally, this tool also asks users to tag an image as Fake or Legitimate. In a similar vein to Buntain and Goldbeck's research, the crowd's knowledge is leveraged to detect fake news. This research introduces a user accessible way to verify fake news in the form of misrepresented photos and images and presents an application of reverse image searching to understand an image's text and temporal context. Further research and improvements to this tool should look more carefully at characteristics of fake users, since there are services available to acquire more followers and generate more posts.

In "Web Video Verification using Contextual Cues" by Papadopoulou et al., the authors propose and demonstrate effectiveness of methods for identifying videos with fake content. The authors observe that User Generated Content (UGC), such as photos and videos, plays a major role in news reporting. Since UGC is unverified, it can lead to the circulation of rumors and fake news. While algorithms have been developed to detect tampering in videos, oftentimes, UGC may be untampered with, but rather falsely described or represented, such as lying about the time and place a video was recorded. The authors developed two distinct classifiers to apply to YouTube videos, one at a comment-level and one at a Video-level. The authors trained the comment-level classifier on The Image Verification Corpus dataset, which contains real and fake tweets from past events.

The authors also assembled an annotated dataset of fake and real videos, the Fake Video Corpus, to train a Video-level classifier. Both classification models had 0.88 precision, and an ideal fusion of these two classifiers could deliver perfect accuracy, though such a fusion has not yet been discovered. The results also show a difference in the distribution of comment credibility estimate distributions between real and fake videos. Real videos have a unimodal distribution of comment credibility, while fake videos have a bimodal distribution. This research targets fake news in the form of video content and develops models that allow for the identification of misrepresented videos, similar to Saez-Trumper's work in identifying misrepresented photos on Twitter. Some shortcomings are the small size of the Image Verification Corpus dataset, which prevents the models from being evaluated more exhaustively and better understood. Additionally, the bimodal nature of the comments on fake videos and the high concentration of relatively credible comments lacks explanation.

Like the work by Papadopoulou et al. where responses to YouTube videos were used to classify content as real or fake, research has also been done on verifying content on Facebook by looking at user response and engagement. In "Some Like it Hoax: Automated Fake News Detection in Social Networks" by Tacchini et al., the authors propose two classification techniques that look at "likes" on a Facebook post to identify hoaxes and misinformation. The research is motivated by the use of social networking websites as vectors for spreading misinformation, and the need for automated methods of assessing post accuracy. The authors consider a dataset of Facebook posts and users, where the posts originate from Facebook pages that deal with either scientific topics, or with conspiracies and fake scientific news. From analyzing the datasets, the authors observed that most posts have few likes, and most users like few things, though some posts and users having a high number of likes. One of the classification techniques the authors propose

is a logistic regression that considers a user's interaction with a post as a feature. The authors also propose an adaptation of Boolean label crowdsourcing techniques. The logistic regression classification technique had accuracy exceeding 99%, and the Boolean label crowdsourcing technique exceeded 99.4% accuracy. Additionally, the models are accurate even with only 1% of posts in the dataset used for training. BLC was more accurate with a smaller training set size. The research contributes two effective and highly accurate means of classifying false posts on Facebook. Additionally, the authors observed patterns regarding the distributions of likes on hoax and legitimate Facebook posts. While these results are promising, further research is necessary to determine if these results also apply to political hoaxes as well, as fake political news has been at the center of public debate.

The literature on the detection of fake news has proposed effective means of detecting specific types of fake news, such as Clickbait, misrepresented photos, and misrepresented videos on platforms such as Facebook, Twitter, and YouTube. The literature demonstrates that there are promising, precise means of automatically detecting fake news based on analysis of contents, media, and user engagement and interaction. The literature mostly provides approaches that have been applied and evaluated on collected data from social networking websites, but lacks proposals for integrating these methods into social networking websites or tools to assist general users in navigating news and content in the era of fake news. The literature is also limited by the types of fake news datasets currently available, preventing broader evaluation of the efficacy of fake news detection methods and models. Further research should attempt training and evaluating detection methods on more broad collections of fake news, and also implementing and architecting Webtools to put fake news detection methods into practice.

Works Cited

Buntain, Cody, and Jennifer Golbeck. "Automatically Identifying Fake News in Popular Twitter Threads." *Smart Cloud (SmartCloud), 2017 IEEE International Conference on*. IEEE, 2017.

Chen, Yimin, Niall J. Conroy, and Victoria L. Rubin. "Misleading online content: Recognizing clickbait as false news." *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. ACM, 2015.

Papadopoulou, Olga, et al. "Web Video Verification using Contextual Cues." *Proceedings of the 2nd International Workshop on Multimedia Forensics and Security*. ACM, 2017.

Saez-Trumper, Diego. "Fake tweet buster: a webtool to identify users promoting fake news ontwitter." *Proceedings of the 25th ACM conference on Hypertext and social media*. ACM, 2014.

Tacchini, Eugenio, et al. "Some like it hoax: Automated fake news detection in social networks." *arXiv preprint arXiv:1704.07506* (2017).