Elizabeth Williams

Professor Stacey Suver

CSE 300

13 April 2018

A Browser Extension for Identifying Fake News and Misinformation on Social Media

Following recent current events such as the outcomes of national elections, fake news and the spread of misinformation on social media has been recognized as a serious issue in the age of social networking websites and user generated content. Social networking websites and the internet contain vast amounts of content, making it far too costly and practical for humans to evaluate the veracity of content. These combined issues have led to the recent emergence of research regarding fake news, including methods for automatically detecting it. The problem is a natural candidate for computational approaches since the data natively exists digitally, and techniques already developed in the fields such as Machine Learning, Social Network Analysis, and Natural Language Processing can be applied to this problem. The research project is to develop a web tool in the form of a Web Browser Extension that identifies fake news content on social networking websites such as Facebook and Twitter by implementing techniques for identifying fake news that have been discussed in the literature. This project will offer a proof of concept for a tool that general users can leverage to verify content and identify misinformation. It will also provide an opportunity for fake news detection methods to be synthesized, applied, and evaluated to true unstructured news feeds instead of just static datasets.

Although research on fake news and methods for detection is an emerging field, there are promising results in accurately classifying fake news of different content types and originating

from different platforms. So far, methods have been proposed for automatically identifying Clickbait, a form of false news, through a hybrid approach using text and non-text cues. Headlines likely to indicate false news can be evaluated by training a classifier to identify the probability of individual words belonging to a prototypical clickbait headline, then using Bayes' formula or vector distances to evaluate an entire headline's likelihood of being clickbait, misleading, or false (Chen et al., 17). A key non-textual cue for identifying Clickbait proposed by the literature was an image incongruent with the headline, which can be identified through behavioral analysis of user interest (Chen et al., 17). This research on identifying clickbait proposes ways that text classifiers and user behavior analysis can be used in identifying false and misleading content. Research has also been done on identifying fake news in popular Twitter threads. Twitter credibility datasets PHEME and CREDBANK were used to train classification models on Twitter conversations and found that these classification models were able to classify accurately most of the time when evaluated against the BuzzFeed Fake News Dataset (Buntain and Goldbeck, 213). This research shows there are reliable methods of detecting fake news on Twitter through low credibility Tweets. Research has also shown that looking at context, such as responses and reactions to social media posts, can be useful in determining content's veracity. Facebook posts could be classified with 99% accuracy by looking at who liked a post (Tacchini et al., 9). A classification model for YouTube videos based on comments had 0.88 accuracy (Papadopoulou, et al., 8). The research I propose will synthesize these fake news detection methods and demonstrate their efficacy in real news feed settings.

The research project will involve the development of a Web Browser Extension as the client for users, which will connect to a server to perform tasks such as classification and also interface with a datastore containing data to be used in the classification and lookup process. The

Web Browser Extension will identify sections of the web page that contain potential fake news content. These sections would include Tweets on a user's Twitter profile or feed, and posts on a Facebook timeline or newsfeed. The extension would be able to identify sections containing content and extract said content through manipulation of the Document Object Model. The extension would use JavaScript to append a button at the top of content sections. Triggering this button would forward text, image, and hyperlink content to the server for further analysis, processing, and classification. The server will implement methods of classifying and identifying fake news and misleading information as described in previous research on automatically detecting fake news. These methods include classification through the credibility of text content in conversations. The classifier on the server will be trained through CREDBANK and PHEME, then continue being trained through new requests to the server and unsupervised learning methods. The server will also implement methods for identifying misrepresented photo content through reverse image searching, then comparing the context and date of the closest match to the post that the image was found in. The datastore will maintain records of image hashes and associated data as reverse image searches are executed. Hyperlinks will be analyzed by the headline and image shown in the preview, and by content through Natural Language Processing techniques. While this analysis will be costlier, the datastore will store the results of classifications on a hyperlink, so that its veracity and attributes can be immediately looked up on future occurrences, as oftentimes a link to online content may go viral. Some limitations of the research methods are that not many user analysis and social network analysis techniques will be used in this tool, due to the high overhead of collecting and analyzing this data per request, and privacy concerns regarding collecting and storing this data in advance. Boundaries of the research are that the Web Browser Extension would be compatible with Facebook and Twitter, but not other social networking websites yet. This is

because the methods proposed thus far in the literature have been applied to Facebook and Twitter effectively, and previous research has been able to apply similar techniques to both social networking websites. Additionally, the CREDBANK and PHEME datasets are in English, so the tool will not be compatible with content in other languages. The research is innovative in that it provides means for users to verify different types of content, especially user generated content, as it appears on their social media feeds. It would be a proof of concept for future tools that could curb the issue of fake news. While these methods have been evaluated already, they have mostly been evaluated on static datasets, and not on dynamic social media feeds. It could lead to future tools to assist users in identifying fake news that does not rely on social media platforms, and the openness of the methods, research, and source code could make it more transparent than fake news detection features that platforms such as Facebook and Twitter may potentially add in the future.

This research would demonstrate a proof of concept for a platform and system architecture for a Fake News Detection Web Browser Extension that can be leveraged by social media users. It will implement methods from the literature that have been either only proposed, or only evaluated on static datasets, onto real content found in any user's social media feed. The efficacy of such methods could then be evaluated on content streams of actual users instead of on preexisting static datasets, and a larger collection of fake news data could be collected. This tool will also synthesize multiple existing methods of detecting fake news onto a single platform, such as combining reverse image searching techniques to identify misrepresented images, along with credibility analysis of the accompanying text to improve the accuracy of identifying false information. Overall, this tool could provide novel and cutting-edge methods of identifying different forms of fake content on social media in a way accessible to users, helping users stay informed and avoid being misled by fake news and avoid spreading fake news.

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.

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