Fake news detection/news credibility ranking

Brian Edmonds - MS Information Security
Xiaojing Ji - MS in Computing Systems
Xingyu Liu - MS in Computer Engineer
Shiyi Li – MS in Computer Science

I. Motivation and Objects

This past presidential election, the American people were overwhelmed with the proliferation of "fake news" articles that altered the narrative (and perhaps the results) of the election. The articles and social media posts featured bombastic headlines and made outrageous claims regarding the candidates. A Buzzfield analysis showed that, despite the mainstream media's best efforts, by election day, more people were "engaging" with Fake News on Facebook than that of traditional news outlets. [1]

Former President Obama discussed this phenomenon in a New Yorker magazine feature following the election of Donald Trump. The article summarized Obama's views:

> "'The new media ecosystem 'means everything is true and nothing is true,' Obama told me later. 'An explanation of climate change from a Nobel Prize-winning physicist looks exactly the same on your Facebook page as the denial of climate change by somebody on the Koch brothers' payroll. And the capacity to disseminate misinformation, wild conspiracy theories, to paint the opposition in wildly negative light without any rebuttal—that has accelerated in ways that much more sharply polarize the electorate and make it very difficult to have a common conversation.'" [2]

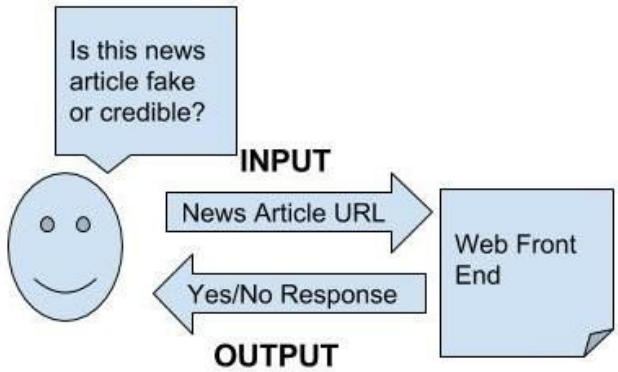
Additionally, the New York Times ran an article that recounted how a Twitter post, written by a user with no formal journalism background, had spread like wild-fire throughout the internet. [3] The post claimed that a busload of anti-Trump protesters were being paid by the Hillary Clinton campaign to attend protest. This post was later debunked, but continued to spread and be retweeted repeatedly.

Given the large-scale implications of a free and fair press in a democracy, the general public needs a more scientific and open approach of discerning baseless news stories from credible, fact-based journalism. We intend to work on a tool that ingests a news article URL and assigns a credibility score to it, indicating whether or not the user should trust it as factual or should conduct further investigation and make their own decision.

II. Proposed Work

We plan to build a web-based application or browser extension to help users identify if a news source is reliable. Our initial definition of reliable and unreliable will rely on the human-curated data http://opensources.co. OpenSources.co is has a list of about 20 credible news websites and a list of over 700 fake news websites. We will begin by building a profile of these sites, crawling both reliable and unreliable sites. The crawled information will be stored in the local machine for further data processing including but not limited URL extraction and author analysis. Additionally, external libraries of some machine learning techniques like LSTM in Recurrent Neural Network(RNN) can be applied for data classification/prediction on the backend server.

The below figure shows the basic user-story that we wish to create. A user will be able to input a news article URL into the application and the application will make a judgement and return a yes or no response as to whether or not the article is credible as journalism.



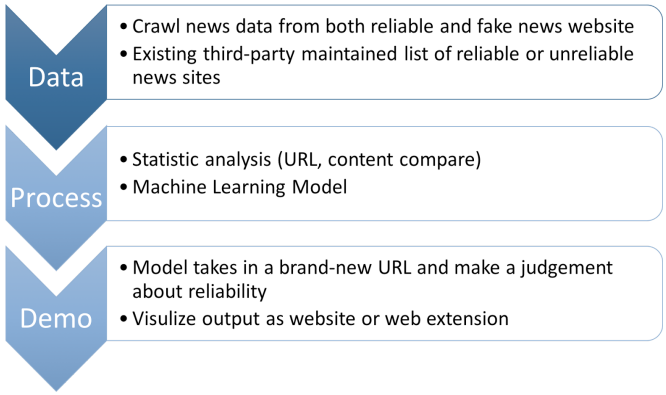
III. Related Work

Recently, major tech companies and computer science students have begun making attempts thwart the spread of fake news on the internet. Google and Facebook both made announcements that they would combat fake news proliferation on their respective platforms. [4] Google appears poised to make policy changes that will eliminate the financial incentives to spread fake news by limiting advertising to deliberately misleading sites. Facebook is following a similar policy route.[7] Additionally, Daniel Sieradski received press coverage for developing a Chrome browser extension that compares the current web page with a list of unreliable websites. [5] Finally, A student from Stanford developed and released a tool that claims to use neural network machine learning techniques in order to identify reliable and unreliable news sources. [6] After it has been released in 2016, the developer didn't uncover implementation details of how it applies neural network in the detection. On HackPrinceton 2016, a group of college students attempted to design a browser extension to detect fake news. Now the project is still under development and it needs future contributions to become functional. [8]

Both Google and Facebook's approaches to fake news reside in the policy actions rather than a technical solution. The BS Detector developed by Sieradski is a useful starting point for us, but relies on humans to curate a list of bad news sites. The Stanford student's solution interests us in that it claims to use Machine Learning to identify how reliable a news site is, but does not publish or reveal any of its methods or algorithms. The community needs an open solution that makes use of statistical analysis and potentially machine learning to identify baseless news articles at network speed. We strive to construct a study of fake news and implement some techniques to find it.

IV. Plan of Action

Below is our high-level plan of action our application:



We plan to collect and mine the following data "features" about a given news article. In some cases we'll try and connect the collected information with other publicly available information to further profile each news article.

TABLE I. Needed Data

Data Point	***Additional Information desired about data point***	***Hypothesis***
Length of Article	Not applicable	None
Author of Article	Does the author have a public profile on Facebook or other social media? What else have they posted?	None
Headline	What is the structure of the headline? Does it contain inflammatory words?	None
Keywords of Article (Frequency analysis)	Do the keywords of an article trend toward reliable or unreliable news sites?	None
How many reliable news sites have linked to this article?	Not applicable	Articles that have been linked to by many different credible organizations will have a strong
How many citations/links does the news article have?	Not applicable	Articles with high number of links to other credible domains or news sources will most likely be credible as well.
Domain/Whois Analysis	How long has the domain name been registered? Who is it registered to?	Domains that have been registered for less than 1 year will most likely not

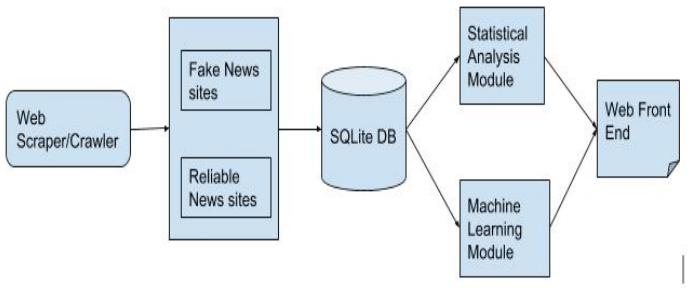
		contain credible news sources.

Libraries and Technology: We plan to use the following libraries and technologies to build a statistical analysis.

- Scrapy - Python library https://scrapy.org/
- News Site Curator: opensources.co
- Platform: Flask 2.0, Chrome
- Database for storing analyzed results (if necessary): SQLite
- Potential Machine Learning Libraries: TensorFlow, scikit-learn, theano, Pattern

V. Implementation Details

The crawled data can be stored locally during the developing phase. The analyzed result for future detection/prediction can be organized in SQLite database. The backend most likely will be implemented on the Flask platform with results stored in a SQLite database. The front end will send information from given webpages to the backend and display the returned credibility results to users. In the above diagram, we demonstrate our plan to collect the data and, after analysis, present it to the end user. We choose a web-based presentation with the hopes that it will enable many users to view our results. The statistical analysis and machine learning module will make the judgements about whether a given article is fake news or not. A user will be able to input a news article URL and the application will return a credibility score about whether it is a reliable news source or not.



Predicted Application Architecture shown above.

TABLE II. TiemLine

Milestone	***Date***
Project Proposal	Feb 6th - Feb 10th
Meeting with professor and revising the proposal based on feedbacks	Feb 11th - Feb 19th
Collect and Clean Data	Feb 20th - Feb 25th
Analyze Data and Try Statistic Analysis	Feb 26th - Mar 10th
Train different machine learning model	Mar 1st - Mar 31st
Data Visualization to Show Result (Web Application or Extension)	Apr 1st - Apr 7th

VI. Evaluation and Testing Method

In order to test if our fake news detection system is successful, we will feed it an URL that we know to contain fake news. If our system is able to give it a low credibility score, than we know that we have succeeded. Another measure of success is if we are able to produce a statistical profile of what data points correlate to a fake news article. We'll be collecting much data on both fake news and credible news and presenting findings and conclusions about this data could be valuable. It is easy for trained humans to identify valid news from fake news, but it may not be so easy for computers. Helping move the community forward in this regard would warrant our project a success.

References

[1] "This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook." *BuzzFeed*. N.p., n.d. Web. 06 Feb. 2017. <https://www.buzzfeed.com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook?utm_term=.prd9xn7b62#.jt7DX9Eowx>.

[2] Remnick, David. "Obama Reckons with a Trump Presidency." *The New Yorker*. The New Yorker, 05 Jan. 2017. Web. 06 Feb. 2017.

[3] Maheshwari, Sapna. "How Fake News Goes Viral: A Case Study." *The New York Times*. N.p., n.d. Web. 6 Feb. 2017. <https://www.nytimes.com/2016/11/20/business/media/how-fake-news-spreads.html>.

[4] Isaac, Mike, Nick Wingfield, and Katie Benner. "Google and Facebook Take Aim at Fake News Sites." *The New York Times*. N.p., 14 Nov. 2016. Web. 6 Feb. 2017.

[5] Ravenscraft, Eric. "B.S. Detector Lets You Know When You're Reading a Fake News Source."*Lifehacker*. Lifehacker.com, 17 Nov. 2016. Web. 06 Feb. 2017.

[6] Dormehl, Luke. "A 19-year-old Stanford Student Has Created a 'Fake News Detector AI'."*Digital Trends*. N.p., 20 Jan. 2017. Web. 06 Feb. 2017. <http://www.digitaltrends.com/cool-tech/fake-news-detector-ai/>.

[7] Love, Julia, and Kristina Cooke. "Google, Facebook Move to Restrict Ads on Fake News Sites." *Reuters*. Thomson Reuters, 15 Nov. 2016. Web. 06 Feb. 2017.

[8] Goel, Anant, Nabanita De, Qinglin Chen, and Mark Craft. "Anantdgoel/HackPrincetonF16."*GitHub*. N.p., 30 Jan. 2017. Web. 06 Feb. 2017. <https://github.com/anantdgoel/HackPrincetonF16>.