In News We Trust

Math 402: Modelling With Uncertainty and Data 1

December 14, 2018

Jane Margaret Cox

Abstract

The issue of rapidly identifying fake news is one of the most important problems facing democracies around the world. The purpose of this project is to first prepare databases of real and fake news articles and, second, use these databases in linguistic-based and network-based statistical approaches to identify probable fake news articles. The linguistic-based approach considers the relationship of politically charged word usage in the main text of articles, while the network-based approach considers the behavior of respective news articles on the social media network of Facebook. In preliminary studies, network-based approaches had greater initial success than linguistic-based one.

Introduction

In the lead up to and in the aftermath of the 2016 election, the term 'Fake News' was used to describe everything from Russian interference in the US election[1] to news stories that disagreed with a person's political opinion[2]. Despite the ubiquity of misleading stories, the identification of fake news remains an elusive goal for government[3] and corporate players alike[4]. The goal of my project is to apply computational uncertainty mathematics to the study of real and 'fake 'news stories using the following linguistic and network parameters: word usage, sentence structure, and social media likes, shares, and comments. For the purposes of this project I use the strictest definition of 'Fake News'; that is, "articles which have no factual basis but are published in the style of news articles to create legitimacy"[5]'. I do not consider articles that are exaggerated or hyper-biased to be fake news.

Fake news articles of this kind (as opposed to satirical or parody articles) are almost always malicious[6]. They were deployed to destabilize Ukraine in the lead up to the Russian Invasion of Crimea (REF). Fake news has been used to cause unrest and polarization in countries across the

---

[1] Prier, Jarred. "Commanding the Trend: Social Media as Information Warfare." *Strategic Studies Quarterly*, vol. 11, no. 4, 2017, pp. 50–85.
[2] Lazer, David MJ, et al. "The science of fake news." *Science.* 359.6380 (2018): 1094-1096.
[3] Scott, Mark, and Melissa Eddy. 2017. "Europe Combats a New Foe of Political Stability: Fake News." nytimes.com
[4] Shu, Kai, et al. "Fake news detection on social media: A data mining perspective." *ACM SIGKDD Explorations Newsletter*19.1 (2017): 22-36.
[5] Tandoc Jr, Edson C., Zheng Wei Lim, and Richard Ling. "Defining "fake news" A typology of scholarly definitions." *Digital Journalism* 6.2 (2018): 137-153.
[6] Tandoc, *Digital Journalism* 2018

world. The exact influence of such articles is hard to quantify, but the possible effects can be chilling[7]. Given the inclusive process of democracy, democratic republics like the United States are particularly vulnerable to polarizing news campaigns based on fallacious material[8]. Authoritarian governments, such as China and Russia, may also use fake news events to manipulate and control their own people.

Despite the urgency of such a task—rapidly assessing whether or not news should be constituted as 'fake news[9]'—comprehensive automated methods for identifying Fake News articles remain elusive[10]. There are currently two main analytical approaches to discovering Fake News: linguistic approaches, which analyze the content of deceptive messages, and network approaches, which analyze how false news stories behave on networks such as social media[11]. Both approaches have their merits and disadvantages. This project will prepare and consider data sets useful for both approaches.

A great challenge to researchers and programmers is the nature of online news data. Current online fake news checkers generally fall into two broad categories[12]: individual story checkers, such as those whose work appears in snopes.com, or sweeping compliations of url's and sources that tend to host fake news stories, such as the online bs-detector extension for chrome. Due to the time-consuming nature of classifying articles, the available data sources are restricted to relatively small databases of articles that are guaranteed to be false or real, curated by researchers, and much larger databases of articles whose classifications are murkier, gathered from trusted and non-trusted news sites. While this approach to data collection results in sufficiently large datasets for analysis (minimum of 10,000 articles), it is problematic. News sources that carry false articles, such as the *RT*, formerly "*Russia Today*", which is funded by the Russian government usually carry truthful articles to boost their credibility. Conversely, highly trusted sites, such as *The Wall Street Journal*, include opinion pieces that may reflect a high level of bias. These aspects complicate the creation of a large-scale dataset without human judgement.

Data

For this project I considered three different data-sources: A database of 12,999 fake news articles[13] (called hereafter the fake news dataset), containing both linguistic and network information, a database of 142,570 real news articles[14] from a variety of sources containing

---

[7] Mickey Robert, Levitsky Steven, and Way, Lucan A. "Is America Still Safe for Democracy". *Foreign Affairs*, 2017

[8] Prier, *Strategic Studies Quarterly*. 2017

[9] Allcott, Hunt, and Matthew Gentzkow. "Social Media and Fake News in the 2016 Election." *The Journal of Economic Perspectives*, vol. 31, no. 2, 2017, pp. 211–235.

[10] Wang, William Yang. "" liar, liar pants on fire": A new benchmark dataset for fake news detection." arXiv preprint arXiv:1705.00648 (2017).

[11] Conroy, Niall J., Victoria L. Rubin, and Yimin Chen. "Automatic deception detection: Methods for finding fake news." *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*. American Society for Information Science, 2015.

[12] Vargo, Chris J., Lei Guo, and Michelle A. Amazeen. "The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016." new media & society 20.5 (2018): 2028-2049.

[13] Risdal, Megan. "Getting Real about Fake News." *Kaggle*, 25 Nov. 2016, www.kaggle.com/mrisdal/fake-news

[14] Thompson, Andrew. "All the News." *Kaggle*, 20 Aug. 2017, www.kaggle.com/snapcrack/all-the-news/discussion.

linguistic information (real news dataset), which I curated to exclude sources that are unanimously considered unduly biased, and a database of 27,384 mixed real and fake news articles[15] (mixed news dataset), containing network information. All data were collected during 2016

The datasets came from contributors on the online dataset database, Kaggle. When building a dataset to be used for classification, it is key that the data are clearly labled and have distinct, correct classifiers. Mis-classifying at this step can result in incorrect models with low accuracy[16]. For instance, if articles that are fake are labeled as real, a statistical or machine learning model would incorporate aspects of those articles into their framework for real articles, even if those aspects are absent in real news articles[17]. As A major concern I had while studying sources involved any biases that the original data collectors may have had, such as collecting Fake News articles exclusively from sources supporting one side of the political spectrum. Due to such possible biases, I used the website mediabiasfactcheck.com to check that there were equal representations in the data. Given the individual biases of writers and reporters, eliminating all biases in the real news dataset is unrealistic; however, sources that had high bias ratings, such as Fox or MSNBC were excluded. While it must also be noted that there is no guarantee that online bias trackers are unbiased themselves, a more objective result can be obtained by comparing perceived bias across several sites.

However, even using such an approach leaves room for fake articles to creep in. There may be a bias in how data were gathered and classified. For example, my dataset of fake news relies on an online fake news application called “The BS Detector”. It relies on OpenSource, an opensource database that contains news providers who are biased, to determine the classification of a news article as fake or real. Because OpenSource is open source, some of the websites are in flux, and it is vulnerable to online trolls or hackers. However examining the site and changes therein shows it to be legitimate at the time the data was collected. Furthermore, repeated random sampling of the data shows that the content of the articles is varied; that is, the fake news dataset does not appear to have a political bias originating from the websites checked by the online detector.

Another potential weakness in the data is the lack of article overlap between datasets. My initial database of 12,999 fake news articles contained both linguistical and network information; however, for real news articles I was only able to find a linguistical dataset. The final dataset I found contained both real and fake news articles, but lacked linguistical data, containing only the relevant network information. While each of the datasets were gathered from around the same time period (2016), between the datasets there was no article replication, meaning that I could not find an article’s linguistic and network information through comparing datasets. Because of

---

[15] Risdal, Megan. “Fact-Checking Facebook Politics Pages.” *Kaggle*, 5 June 2017, www.kaggle.com/mrisdal/fact-checking-facebook-politics-pages.

[16] Gelman, Andrew, and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[17] Murphy, Kevin P. *Machine Learning: A Probabilistic Perspective*. Cambridge, Mass: MIT Press, 2012. Internet resource

this, I used feature engineering and data analysis to build two separate datasets for linguistical and network analysis.

The final potential issue with the data is it's age. Collected during a highly charged political period in the United States, the 2016 presidential election, both real and fake news articles center around political issues unique to that time. Given how news changes, the content matter of current fake news articles differ from those collected. There are less articles, for example, about Hillary Clinton currently than there were during the lead up to the 2016 election. I considered deleting all articles with reference to Hillary Clinton from the database, but decided not to, due to the fact that a random sampling of the articles showed similar naming conventions to fake news stories today. i.e. "You won't BELIEVE what Clinton just did[18]". Simply swap out the name Clinton for whatever politician of the party the article is attacking. For this reason I chose to leave references to Clinton in; though the content may change, the format remains, for the most part, the same.

Methods: Data Preparation

The Fake News dataset contains articles sourced from over 20 different verified fake news sources. Although fairly clean, the dataset comes with a few issues. Firstly, it includes articles in Russian and Chinese as well as English. Although Russia[19] and China[20] both have relevant misinformation campaign, I limit this project to the analysis of articles written in English, and I excluded all non-English articles. The second issue is that article titles are split between two columns, Title and Thread-Title. After determining that a title was either in one column or another, but never both, I combine the article titles into one column, dropping all articles that do not have a title.

In order to prepare the dataset for linguistical comparisons between the real and fake news datasets, data must also redacted. As the real news dataset does not include any network information, this information is excluded, and column names were normalized between the two datasets.

Finally, this dataset specified the specific type of Fake News each article was, based on the online classifier the bs-detector. Although interesting, the goal of this project is to compare real and fake news, rather than to to compare fake news sources. For this reason, I replace the determiners in this column with 'fake'. The final fake news data frame containing linguistical information has the following parameters:

	author	date	title	text	publication	type
0	Barracuda Brigade	2016-10-26T21:41:00.000+03:00	Muslims BUSTED: They Stole Millions In Gov't B...	Print They should pay all the back all the mon...	100percentfedup.com	fake

In addition to studying how Fake News in this dataset behaved, I also want to consider the network relationships between the articles. For this reason, I additionally prepared the dataset in

[18] Risdal, *Kaggle*. 2017
[19] Prier, *Strategic Studies Quarterly*. 2017
[20] Rawnsley, Gary D. "Old Wine in New Bottles: China-Taiwan Computer-Based 'Information Warfare' and Propaganda." International Affairs (Royal Institute of International Affairs 1944-), vol. 81, no. 5, 2005, pp. 1061–1078

such a way to allow for network analysis. That is, I preserved the network related data, or the Facebook likes, shares, and comments to allow for an in-depth statistical consideration of fake

	author	published	title	text	language	site_url	country	likes	comments	shares	type
0	Barracuda Brigade	2016-10-26T21:41:00.000+03:00	Muslims BUSTED: They Stole Millions In Gov't B...	Print They should pay all the back all the mon...	english	100percentfedup.com	US	0	0	0	bias

news articles' online behavior. The parameters for this dataset are as follows:

The real news dataset contains articles that come from reputable sources, scraped between July 2016 and July 2016. There are 15 news sources included in this dataset: The New York Times, Breitbart, CNN, The Business Insider, The Atlantic, Fox News, The Talking Points Memo, Buzzfeed News, The National Review, The New York Post, The Guardian, NPR, Reuters, Vox, and The Washington Post[21].

The data here present an interesting challenge. While all of the sources are credible, some are either blatantly biased towards one side of the political spectrum (such as Vox or Fox New), while others rate suspicious on the online BS meter that I use in my Fake News source (such as Breitbart). I want the articles in my clean section to be certified completely clean; that is, to have little or no suspected bias. The BS detector has no rating for Buzzfeed, Talking Points Memo, National Review, New York Post, Business Insider, or the Atlantic, either as credible or not credible, so I removed these sources from this list. I recognize that an unintentional bias may result from this.

This still leaves sources that are biased one way or the other. The goal of this project is not to measure biases of various papers- it is to distinguish between real and fake news. Unfortunately, an extremely biased article on either side of the issue can read as a fake news article[22]. Given this, using online bias indicators, I further narrowed my choices of news sources to the following: The New York Times, NPR, Reuters, and the Washington Post. This is not a perfect solution, given that each of these sources do have political biases. However, something that has become abundantly clear as I sift through news articles is that everything is biased; these are just some of the sources that are consistently less biased than others contained in the dataset.

The final cleaning required for the real news dataset was normalizing the column names to match the fake news dataset for linguistical analysis, and adding a column containing the classification of the articles as real news.

To finish creating a clean dataset of both real and fake news articles for linguistical analysis, I merge the fake news and real news linguistical datasets. I added the Breitbart News articles found in the real news dataset to the fake news dataset before doing this. The final combined dataset has the following parameters:

	author	date	publication	text	title	type
0	Carl Hulse	2016-12-31	New York Times	WASHINGTON — Congressional Republicans have...	House Republicans Fret About Winning Their Hea...	real

[21] Thompson, *Kaggle* 2017.
[22] Lewis, David D., and Marc Ringuette. "A comparison of two learning algorithms for text categorization." Third annual symposium on document analysis and information retrieval. Vol. 33. 1994.

The mixed dataset contains a mixture of categorical and numerical data. Categorically, it contains the type of post made on social media: a video, link, text, or photo; the source of the post: mainstream, left, or right; the rating given to the source by a viewer: no factual content, mostly true, mixture of true and false, or mostly false; and the page responsible for the post, which draws from 3 mainstream, 3 left, and 3 right sources. For numerical data it contains the number of shares, the number of reactions, including both likes and dislikes, and the number of comments.

For ease of use, I disregard the account id, post id, the url, and if there the existence of debate in the comment section. I also dropped all rows containing Nans (missing information) and rename some of the columns. I now have three datasets: A clean for linguistic analysis containing both real and fake news articles; a dataset for network analysis of the fake news articles contained in the previous dataset; and a dataset for network analysis of both real and fake news articles.

Feature Engineering

In order to more fully analyze the information in the datasets, I use feature engineering[23] to extract further data. I focus on the categorical linguistic data that can be extracted from my constructed dataset for linguistical analysis. I determine the top n-grams contained in the titles for this data. I also used one hot encoding to prepare the type column (containing if the news is real or fake) for further analysis, and created a column containing the number of Clinton or Trump mentions in an article.

A common approach to textual data is the bag-of-words approach[24]. Bag-of-words considers the n-grams, or the nth number of words in a row and how often they occur[25], then uses their frequencies to classify texts. An n-gram is a continuous sequence of n words that occurs in a given sample of speech or text[26]. It can be letters or even phonemes, but the purposes of this project I considered whole words. For example, if I were to split the sentence “Euclid had a sad dog” into bi-grams, or pairs of words, the data would be as follows: “Euclid had; had a; a sad; sad dog”. While this example is trivial, in the text of a longer article the frequency of “sad dog” may be much greater than “Euclid had”, potentially affecting the classification of that article.

I first studied the most common n-grams in fake news texts, then the real news text. In part due to the nature of the English Language, bi-grams do not yield many phrases of significance, even after clearing the data of filler words such as ‘the’ and ‘and’. Tri-grams, or triples of words, offer a little more insight, but the most interesting contrasts occur with 4-grams. Unfortunately, there are no obvious n-grams present in the fake news article text data that are also absent in the real news article text data. That is, for the data that I have, bag-of-words does not seem to be an effective classifying technique. It is however, able to demonstrate the general behavior of both fake and real news articles.

---

[23] Murphy, *Cambridge University Press*. 2012

[24] Zhang, Yin, Rong Jin, and Zhi-Hua Zhou. "Understanding bag-of-words model: a statistical framework." International Journal of Machine Learning and Cybernetics 1.1-4 (2010): 43-52.

[25] Goodfellow, Ian, et al. Deep learning. Vol. 1. Cambridge: MIT press, 2016.

[26] Goodfellow, *Cambridge: MIT press*. 2016.

The most common 4grams for both fake and real news is the United States, with phrases containing 'United States' or 'The US' occurring in 18.5% of fake news articles and 22.5% of real news articles. This suggests that subject matter is not a good choice to differentiate real and fake news. This follows common logic; an effective fake news article is one that addresses issues concurrently with the real news articles.

Of more interest is the reference to the *New York Times* found in both real and fake news articles. Combining both spellings, the phrase *'The New York Times'* is the third most common 4gram for Fake News, occurring in 5.4% of the fake news articles. It is the fifth most common 4gram for real news, occurring the included two references to 'The New York Times', with it also occurring in 5.4% of the real news articles.

Considering real news, given that articles from the *New York Times* make up roughly a quarter of the articles, it is perhaps understandable that references to this source are common, although further text analysis is needed to verify this. Harder to explain are the phrases prevalent in fake news articles. In a conversation with journalist Elizabeth Maki[27], she explained that one of the largest indicators journalists use to analyze an article's veracity is the number of independent sources quoted in it. Perhaps by referencing a trusted source such as the *Times*, fake news articles seek to increase their credibility. However, this is not the only possibility. The fact that no other news source occurs in any n-grams could mean that the fake news sources are attacking the credibility of *The New York Times*. Further analysis of the texts in question is needed to determine this.

In order to deal with the categorical nature of the fake and real news article type, I use one hot encoding, which prepares the data for numerical analysis. One-hot encoding turns categorical data, or data that is split into exclusive and unique categories (such as real or fake) into numerical data. It does this by adding the same number of columns as there are categories. For each data point, a 0 or a 1 is entered into the added columns. A 1 indicates that the point has that category, while a 0 indicates that it does not. This can be seen in the dataset description below:

PICTURE

As mentioned previously, the ratings were done by source, not individual articles, so this is by no means a perfect description.

The final feature engineering I perform on my linguistic approach data considers the number of mentions of prominent politicians when the data were collected in 2016: Donald Trump and Hillary Clinton. This is to better understand what, if any, trends fake news and real news articles followed in their content.

Results

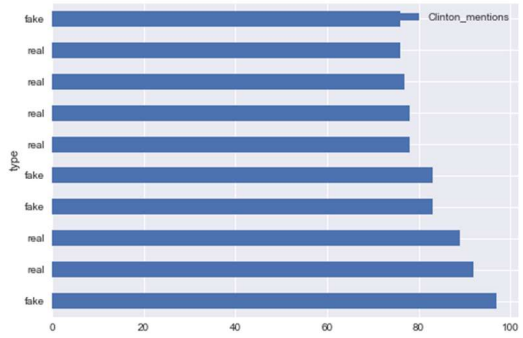
Given the sprawling nature of the problem of identifying fake and real news articles, I pursue three overarching realms of analysis. I initially use a linguistics approach on the real and fake news dataset containing article content, focusing on Trump and Clinton mentions. I then consider

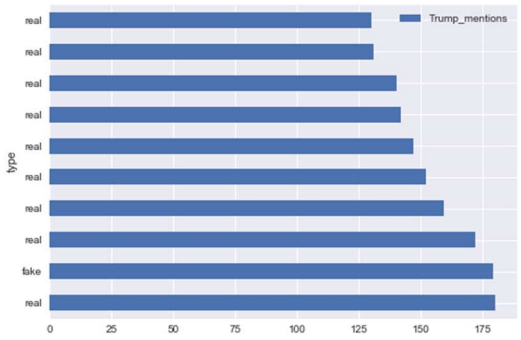
[27] Maki, Elizabeth. Personal Interview. December 1, 2018.

the network behavior of fake news articles contained in that dataset, and I finally consider the network behavior of both fake and real news articles using the mixed dataset.

The initial linguistic based analysis considers the frequency of mentions of the two most prominent people in 2016: Hillary Clinton and Donald Trump. It should be noted that news articles collected in 2018 will likely have a different main focus; however, the goal of this analysis is to consider if either fake or real news articles focused more on the candidates during the election. The 2016 election was one of the most polarizing in our nations history, and this analysis aims to consider which source of news more often referenced the two polarizing people at its epicenter, Clinton and Trump.

The following graphs show the news articles that mentioned Clinton or Trump the most, according to source





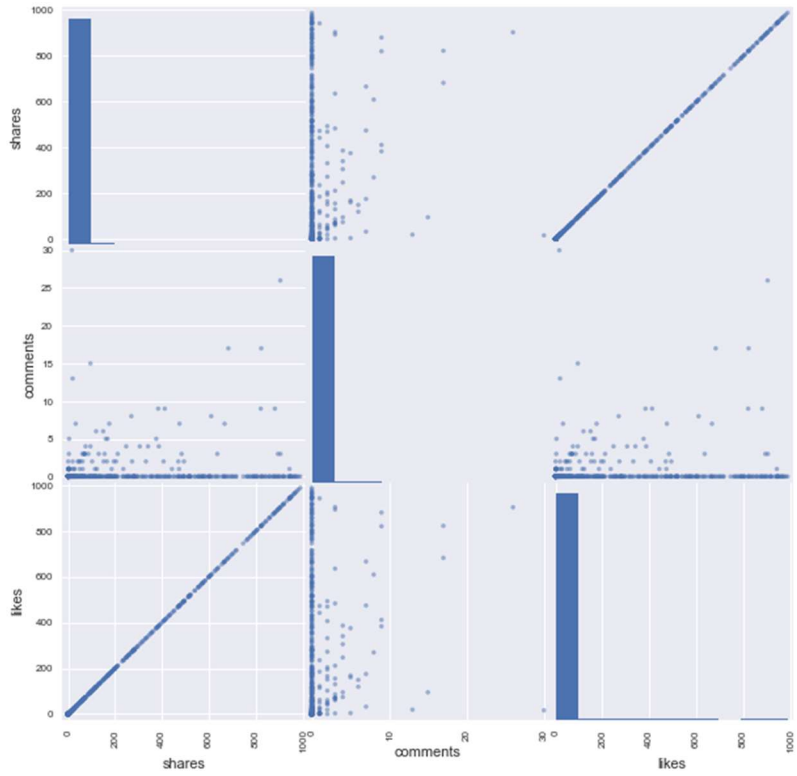
Surprisingly, the articles containing mentions of Trump and Clinton overwhelmingly were written by reputable news sources. I expected fake news articles to have more, which they did not.

It is also interesting to note that, comparing the top mentions in the fake news dataset, there were more fake news articles that mentioned Clinton than that mentioned Trump. This may be indicative of a greater volume of articles attacking Clinton, but I hesitate to draw that conclusion on these data alone. A significant amount of the fake news article came from Breitbart News, an extreme far right online newspaper that ran many articles attacking Clinton during the election. This could have skewed the results.

Finally, note that the top ten are not divided by sentiment; that is, this is not the top 10 anti-Clinton articles or anti-trump articles, just the articles that mention Trump the most. Looking at article titles and texts, the top ten include a good mix of both pro and against articles. From this data and the n-gram analysis considered in the feature engineering, I conclude that basic linguistic methods of analysis are insufficient to differentiate between real and fake news.

In order to better understand the relationship between the various social media presences that a fake news story can take, I consider the network information contained in the initial fake news dataset. I do this analysis separately than the joint analysis because the data here also contains the data that was used in the linguist analysis above, and I wanted to better understand that data

in a network way. I consider the relationships between Facebook likes, shares and comments in an article, as shown in the following scatter matrix:



The relationship between an article being liked and being shared appears highly linear. This suggests that if an article was liked, the person who liked it also shared it immediately, which is extremely unlikely.

To further explore the relationship between likes and shares I perform a linear regression.

For my independent variable I take likes, and for my dependent variable I take shares. The R-squared score for the model is 1.0, meaning that the residual sum of the squares is 0, or in other words that the data arw perfectly linear. Furthermore, the standard error is close to 0 at 2e-18. Such a perfect correlation does not exist for regular social media usage. This strongly suggests that the articles shared here were all promoted by online social media bots. Given what has come out about bots on social media, this may be the result of bots liking and then sharing an article automatically. At first glance, there isn't much of a relationship between either comments and likes or comments and shares. However, after accounting for outliers, it's clear that almost every article has one comment, regardless of how many likes or shares it got. Again, the uniformity here is really weird and appears to be synthetic in nature. Thus any article that behaves like this on a social network is almost guaranteed to be fake.

Dep. Variable:	shares	**R-squared:**	1.000
Model:	OLS	**Adj. R-squared:**	1.000
Method:	Least Squares	**F-statistic:**	2.251e+35
Date:	Thu, 13 Dec 2018	**Prob (F-statistic):**	0.00
Time:	21:23:46	**Log-Likelihood:**	3.7415e+05
No. Observations:	12391	**AIC:**	-7.483e+05
Df Residuals:	12390	**BIC:**	-7.483e+05
Df Model:	1		
Covariance Type:	nonrobust		

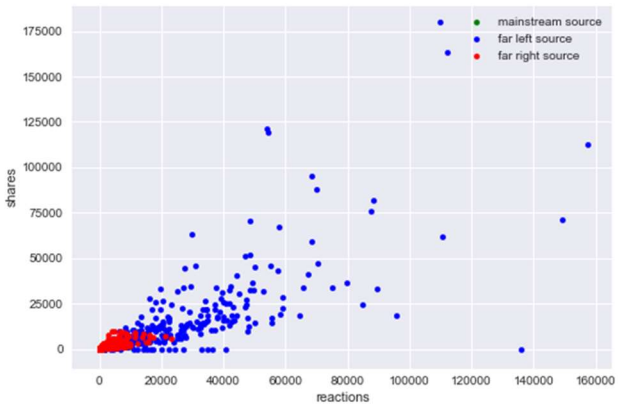
	coef	std err	t	P>|t|	[0.025	0.975]
likes	1.0000	2.11e-18	4.74e+17	0.000	1.000	1.000

Omnibus:	18924.907	**Durbin-Watson:**	1.411
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	6021954.011
Skew:	-9.866	**Prob(JB):**	0.00
Kurtosis:	109.182	**Cond. No.**	1.00

Lastly, I considered the distribution of likes per article. It is highly irregular and does not follow any common distribution, but is useful in understanding how fake news articles behave. Their likes are drastic; that is, either

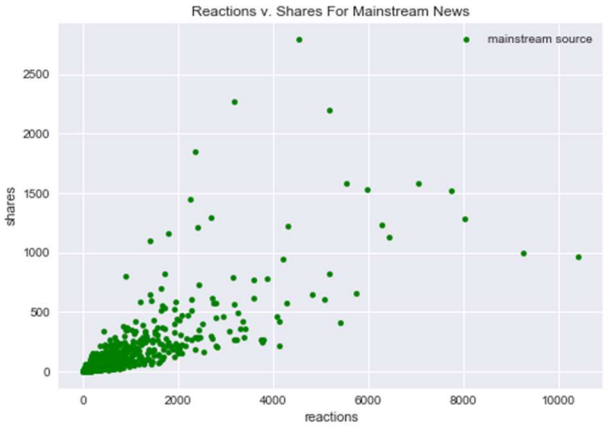
an article gained shares of upwards of 600, or none whatsoever. The articles that became popular don't seem to follow any trends.

In addition to studying the network behavior of fake news, I also considered the network behavior of both real and fake news in terms of reactions (likes and dislikes) and shares on Facebook, according to mainstream, far left, or far right articles.

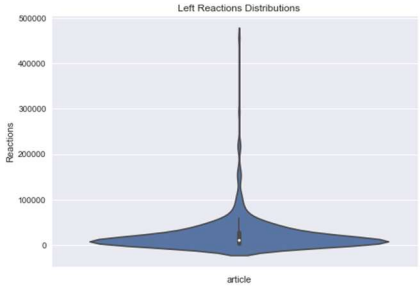


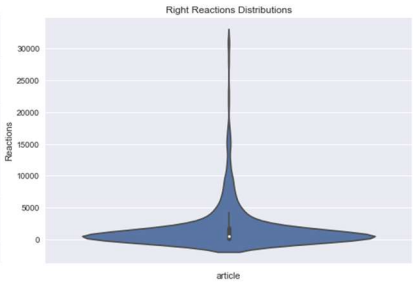
As seen on the graph left, fake news articles propagated by far left sources tend to receive far more attention on social media. Fake news articles from far left sources again receive greater attention than real news articles from mainstream sources. This suggests that if a news story is achieving great network success on a social network it should immediately be considered suspect.

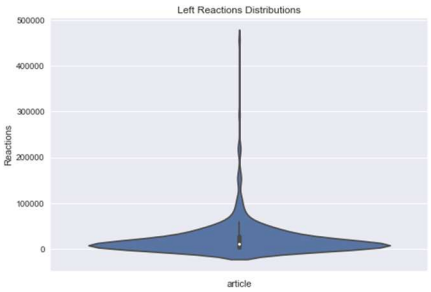
Network behavior for mainstream news sources is much more circumspect, with even the most popular articles obtaining over 50,000 less reactions that those of fake news articles, as seen in the figure right.



The distribution for news articles from all sources resembles a gamma distribution. This matches what is known about items shared on social media; that is, a few items will go viral, but the majority will remain within a relatively small network of like-minded people. These trends are visible when considering the violin graphs of the data:







This matches what is known about items shared on social media; that is, a few items will go viral, but the majority will remain within a relatively small network of like-minded people.

Comparing these results with the network analysis of the fake news sources suggests that, unless the news article in question exists solely in a bot-net, distributive patterns of likes, shares, and comments are insufficient to differentiate between fake and real news articles

Identifying and debunking fake news stories is one of the most pressing challenges for journalists and computer scientists today. It is also one of the most difficult, as shown by this analysis. There are, however, future avenues for research. First, although network analysis proved insufficient, this approach did not consider the nature of the individual players on the network. That is, did the same individuals post news articles often? If so, what sources were they more likely to come from? Second, if initial linguistic analysis failed, the data are prime candidates for deep learning NLP methods. Finally, in order to truly attack this problem, a hand-curated database of news containing linguistic and network data of both real and fake news articles is required. Basing veracity off of source is an insufficiently precise measurement of an article's credibility, and creates datasets of dubious quality. Further research should use such a dataset, and pursue higher methods of analysis in order to truly identify real and fake news.