Exploratory Analysis on Fake News and Unreliable Sources

Report and Analysis By: Andrew Boe

12/14/2017

Abstract

Recently, the phrase “fake news” has infiltrated the political climate. News media and the public have expressed concern about the political ramifications that fake news may have on the world. This project explores a public dataset from kaggle.com that contains about 13,000 “fake news” posts from October 26th, 2016 to November 26th, 2016. It first analyses the timeline and frequency of the posts to discover a trend related to the 2016 United States Presidential Election. It then analyzes the context from the text within by looking for frequent keywords and classifying sentiment using training data from a research study on emotions.

Materials

The fake news dataset used was a file from kaggle.com and was created by a user Megan Risdal. The file was in the form of a .csv and the data contained text and urls for fake news articles from 244 websites that were deemed fake by a web extension B.S. detector. It also had a decent amount of metadata, most of which was not useful to analysis. The data that was used was author, published date, title, text, site url, and type. It was parsed into a list of dictionaries.

The type metadata was taken from B.S. detector which rooted from OpenSources (a list of unreliable sources). Most of the data (about 11,000 of 13,000) was missing a type and given ‘bs’ as a replacement. OpenSources had a publicly accessible github that contained sources.json, a file full of unreliable websites that had a type associated with each. To fill in this missing ‘bs’ data, the list of dictionaries was looped and if the url matched one of the unreliable urls in the .json file, it was replaced with the appropriate type (`fakenewsanalysis.pynb, cell 2, def fix_unknown_types`).

Most of the data used was not numeric, which made it difficult to detect for any bad values. A lot of data was missing, but no list wise deletion was done because the data that was needed was mostly the dates of each published article. The text and titles were cleaned separately by removing punctuation, and setting to lowercase.

A databank called the International Survey on Emotion Antecedents and Reactions (ISEAR) was used to train the classifier used in sentiment analysis. This data came in the form isear_databank.mdb and was converted using Microsoft access into a csv file called isear.csv. It was then parsed into a list of lists containing the training data.

Results

The arrival of “fake news” as a buzzword is often associated with the beginning of the highly controversial United States Presidential Election of 2016. The former republican candidate and current president Donald Trump popularized the phrase and frequently uses it to oust news that he considers biased or unfair. This project analyzes this phrase, but in a different context. It explores news articles that have been flagged as unreliable for either containing sources to, or being a known unreliable source.

The phrase "fake news" is not a new concept, it has been used to describe news containing biased or false information as early as 1575. (Merriam-Webster, 2017). The significant rise in concern regarding fake news however, is new. The spread of fake news in the modern age is mostly attributed to the internet, specifically social media websites like Facebook and Twitter. A survey by the Pew Research Center in 2017 claims that "As of August 2017, two-thirds (67%) of Americans report that they get at least some of their news on social media" (Elisa Shearer, Jeffrey Gottfried; News Use Across Social Media Platforms, 2017). This is concerning, especially when considering the amount of time that users spend on a webpage. Tony Haile, the CEO of Chartbeat, a data analytics company out of New York City studied "Two Billion visits across the web over the course of a month and found that most people who click don't read. In fact, a stunning 55% spent fewer than 15 seconds actively on a page" (Burkhardt, Joanna; Library Technology Reports, Nov/Dec 2017).

There is also a public concern for the legitimacy of the news found in social media. After polling around 70,000 respondents from around the globe in 2017, The Reuters Institute Digital News Report found that only 24% percent of these respondents believe that social media does a good job at separating fact from fiction. Clearly, there is doubt in the news received over social media.

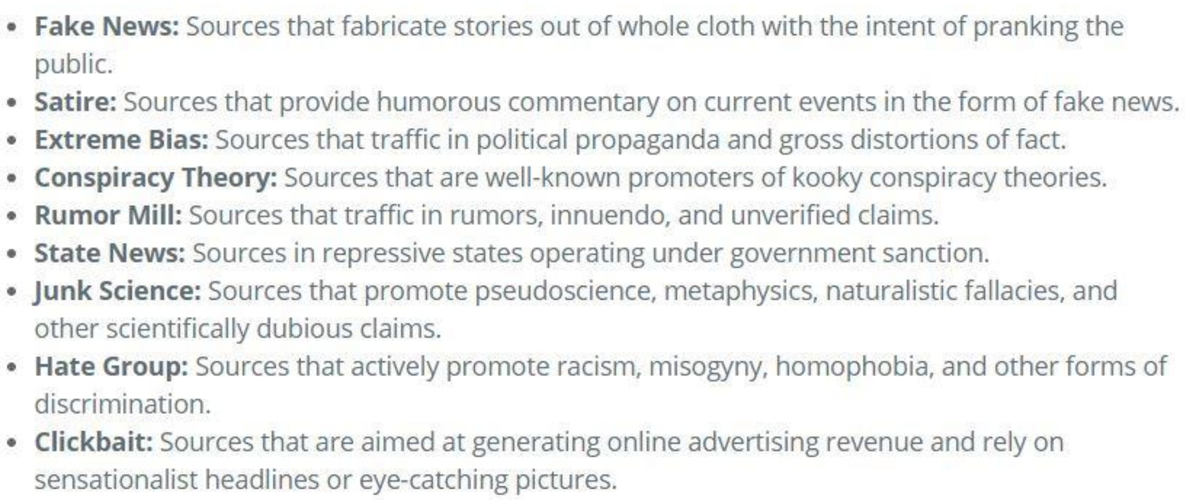
This emphasizes the importance of analyzing the content contained in articles confirmed to be fake or unreliable. It also emphasizes the importance of analyzing how fake news can impact the events of an event like the 2016 Presidential election.

Exploration

This exploration began with a public dataset from kaggle.com of about 13,000 posts from websites flagged as fake news by the browser extension B.S. Detector. B.S. Detector scrapes webpages and looks for links, comparing those links to a professionally curated list of unreliable sources (OpenSources), and determines if that article is legitimate. The articles are published from October 26th, 2016 to November 26th.

The articles contain a type tag that tells more about the specific type of fake news. Figure 1 shows the different types and descriptions (taken from bsdetector.tech). Plots were then created for each type using the matplotlib library in python. The number of bins was set to 30 to accurately represent the 30-day span of the data (`fakenewsanalysis.pynb, cell 4)`. Figure 2 shows the histogram for the all types and figures 3-10 show the subplots of individual fake news types.

There is a clear trend in figure 1, the number of fake news articles appears to be decreasing as time passes on. This leads to two possible conclusions about this dataset. The much more likely conclusion is that there is some bias in the data. The creator of the dataset may have done a significant amount of scraping in the first few days, collecting many articles. Then as time passed on, collected a smaller amount. The less likely scenario is that if the data is legitimate or unbiased, then there may have been some sort of event that causes this decrease in articles. The articles were collected about 15 days before and 15 days after the 2016 Presidential Election, which took place on November 9th, 2016. This could suggest that fake news creators slowed down the amount of news as the election approached. Potentially because the impact that creating more new articles would have negligible affects on the election.



- **Fake News:** Sources that fabricate stories out of whole cloth with the intent of pranking the public.
- **Satire:** Sources that provide humorous commentary on current events in the form of fake news.
- **Extreme Bias:** Sources that traffic in political propaganda and gross distortions of fact.
- **Conspiracy Theory:** Sources that are well-known promoters of kooky conspiracy theories.
- **Rumor Mill:** Sources that traffic in rumors, innuendo, and unverified claims.
- **State News:** Sources in repressive states operating under government sanction.
- **Junk Science:** Sources that promote pseudoscience, metaphysics, naturalistic fallacies, and other scientifically dubious claims.
- **Hate Group:** Sources that actively promote racism, misogyny, homophobia, and other forms of discrimination.
- **Clickbait:** Sources that are aimed at generating online advertising revenue and rely on sensationalist headlines or eye-catching pictures.

Figure 1: Fake News Types in the Dataset

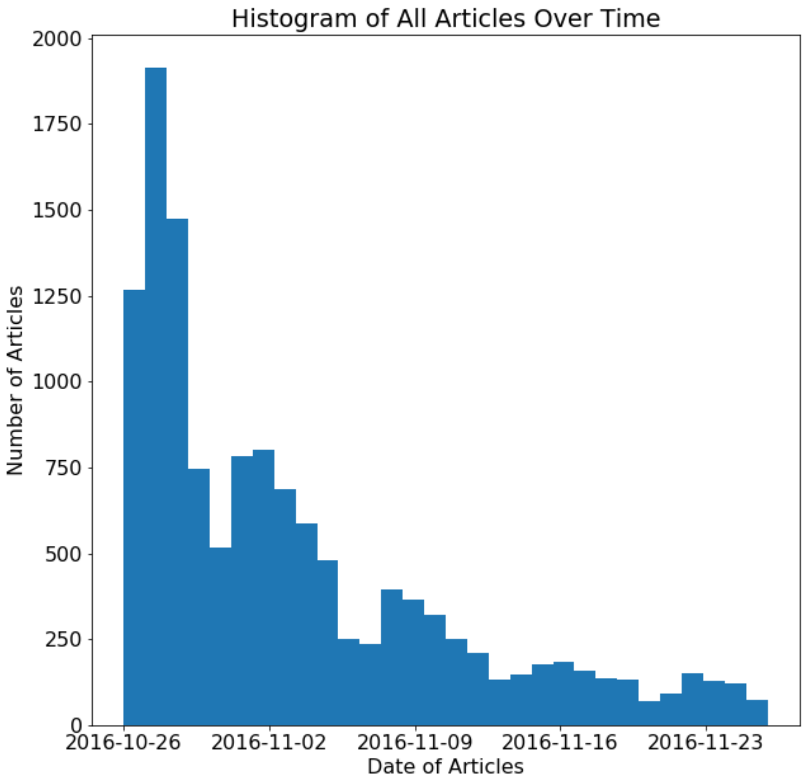
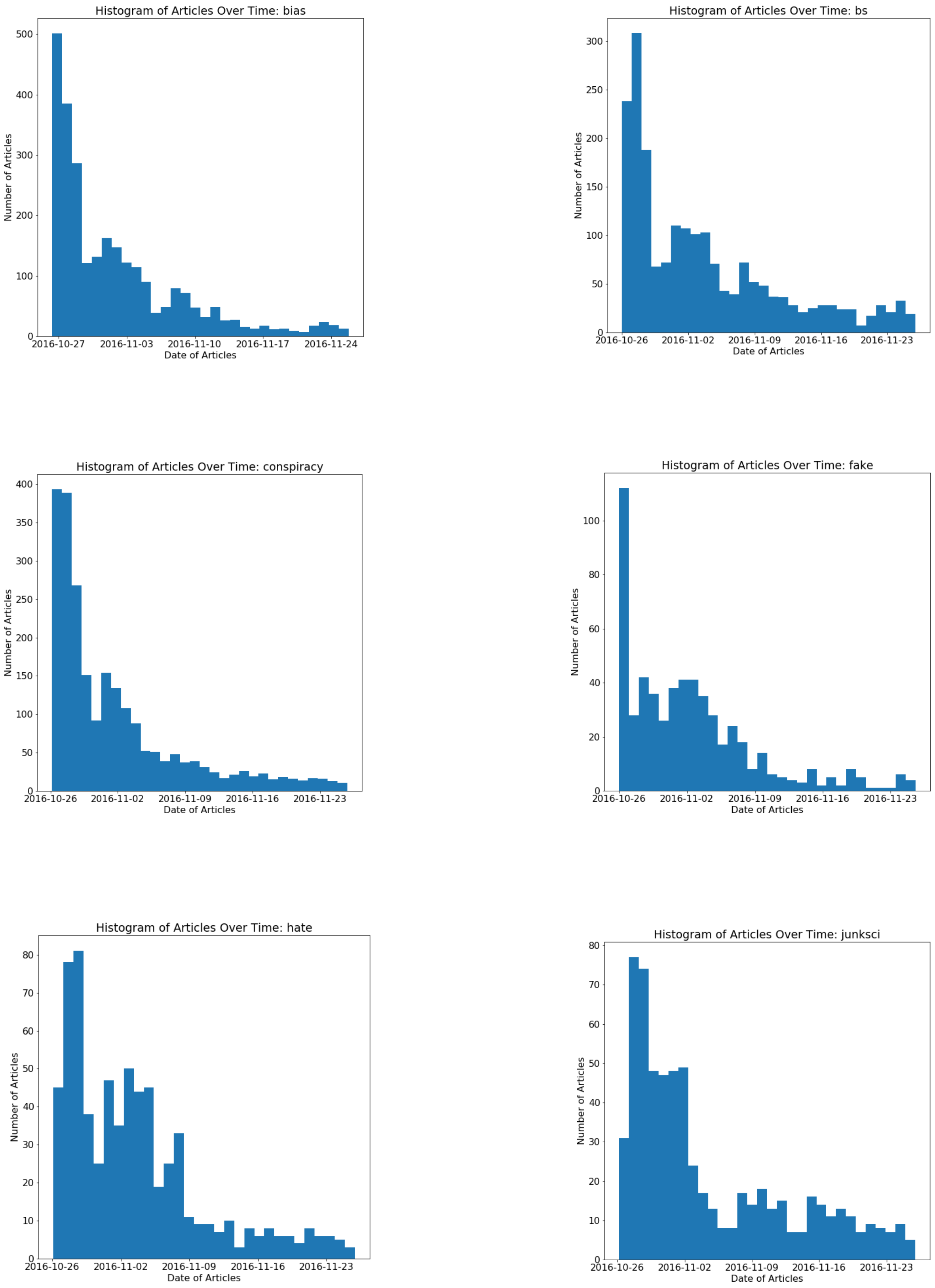


Figure 2: Histogram of All Article Types Over Time



Histogram of Articles Over Time: bias

Histogram of Articles Over Time: bs

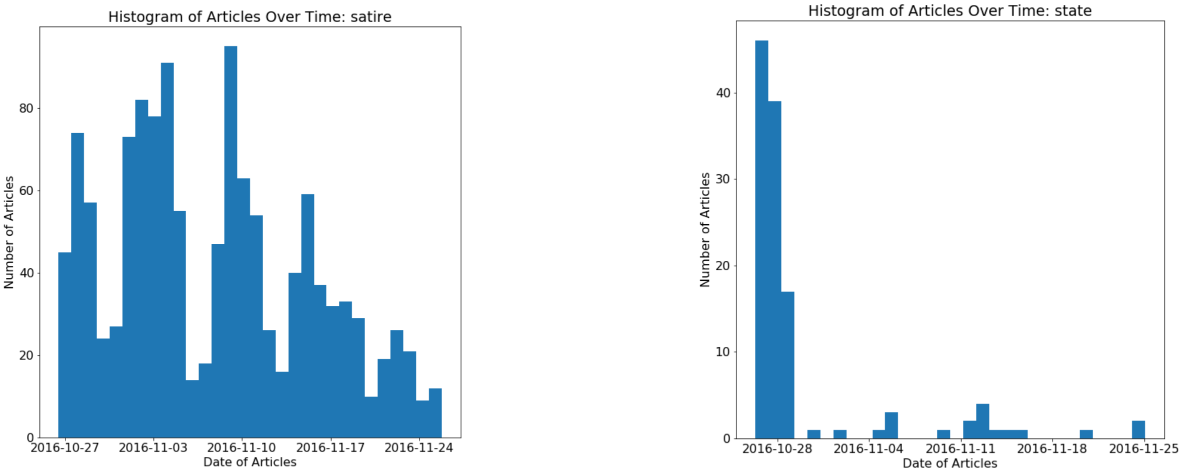
Histogram of Articles Over Time: conspiracy

Histogram of Articles Over Time: fake

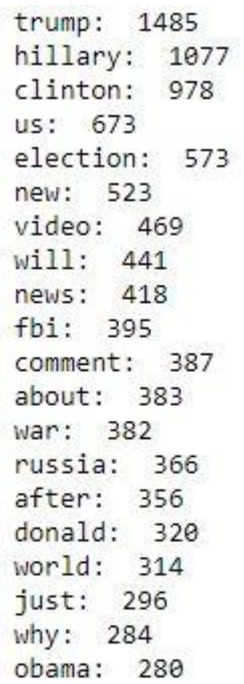
Histogram of Articles Over Time: hate

Histogram of Articles Over Time: junksci

Figures 4-8: Histograms of Dates by Type



Figures 9-10: Histogram of Dates by Type



```
trump:  1485
hillary:  1077
clinton:  978
us:  673
election:  573
new:  523
video:  469
will:  441
news:  418
fbi:  395
comment:  387
about:  383
war:  382
russia:  366
after:  356
donald:  320
world:  314
just:  296
why:  284
obama:  280
```

Figure 11: Top Words in Titles

The subplots suggest, however, that the data is in fact biased. Looking at figure 7, which represents the Junk Science articles (junksci). Although the number of articles is much less than the total, the decreasing trend still exists. Figure 1 describes junk science as “Sources that promote pseudoscience, metaphysics, naturalistic fallacies, and other scientifically dubious claims”. This would suggest that junk science articles are generally not affiliated politically. If there was no bias, then there would probably be no significant downward trend.

There is however, evidence in the text and titles of the articles, that they are mostly politically driven. Figure 11 shows the top 20 words the appear in titles along with the number of times they appear. The top two words after removing a few pronouns, prepositions, and conjunctions, are Trump and Hillary, followed by Clinton, U.S. (potentially us) and election. This implies that most articles have a political motivation and likely have a desired agenda on a reader who might take the article as fact.

Sentiment and Clickbait

The statistic that 55% of people spend under 15 seconds on a webpage implies that not a lot of information is absorbed from the site itself. This prompts exploration into how fake news is designed and how it attracts readers. The type of online journalism that is designed to attract readers with a headline is known as clickbait and is often “a major contributor to the spread of fake news on the internet” (Chen, Conroy and Rubin, 2015).

The fake news dataset is filled with articles that contain titles about current political issues. Most of these articles are likely to contain some sort of clickbait headline. To analyze these headlines, Multinomial Naïve Bayes was used to classify articles as a certain sentiment or emotion. The training data used was from a data set called ISEAR (International Survey on Emotion Antecedents and Reactions). It contains sentences that describe a specific emotion felt at the time. The emotions were then counted to see if there was a majority. Figure 12 shows the list of each emotion and the percent of each classified type for the title of the article. Figure 13 shows the list of each emotion and the percent of each classified type for the text of the article.

This sentiment analysis of titles and text of each post is somewhat conclusive. The majority of article text and titles were classified to anger, 40% and 22% respectively. The next two highest classifications are disgust and fear for both article text and titles. The titles of each article were short, so it is likely that the analysis is less accurate than that of the text of each post. This analysis results in a conclusion that most of the article sentiment is negative. Articles tend to lean toward anger, disgust, and fear. This implies that the agenda of fake news is to defame character, rather than promote a political affiliation.

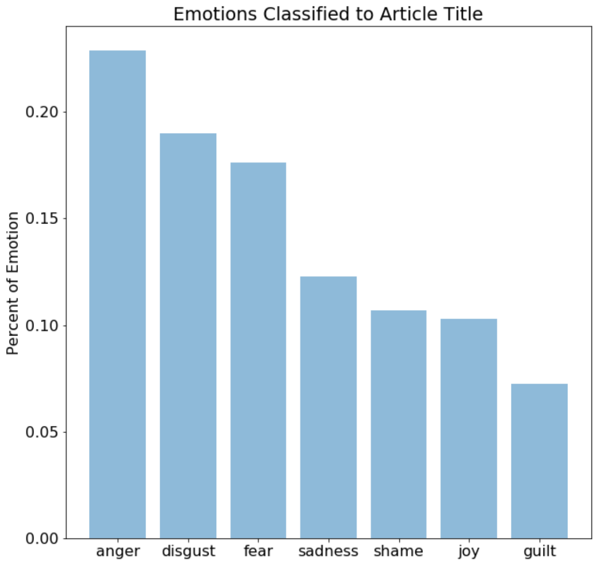


Figure 12: Emotions Classified to Title

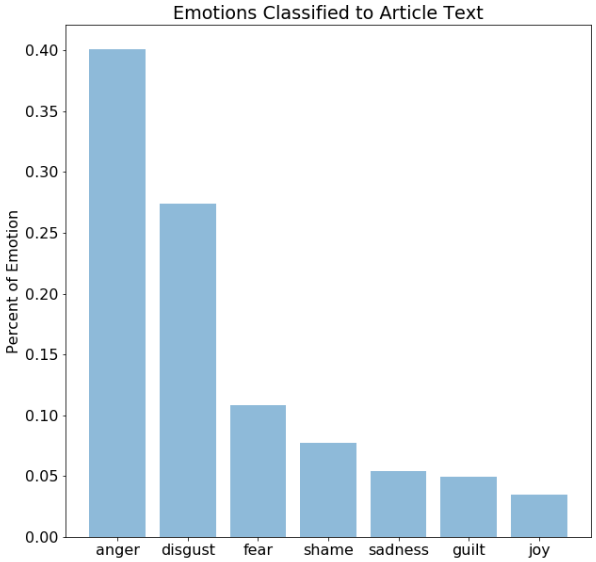


Figure 13: Emotions Classified to Text

Conclusion

The initial analysis of the timestamps of each post shows that a definite conclusion regarding this data is difficult. This is due to a high likelihood of bias. The data, being one of the more easily accessible datasets on fake news, may not be conclusive of any trends during the 2016 election. This means that if fake news was to be properly analyzed in the frame of the election, there would have to be a new dataset that fairly chooses articles based on the actual frequency of published articles during the election. This would be an extremely difficult task as one would have to, without bias, choose websites, and then categorize the news as fake. The B.S. Detector is inherently bias because it bases its decision from a list of claimed unreliable news sources. It is clear, however that there is a target demographic for these fake articles. The large amount words associated with the presidential election in article titles implies that most of the fake news articles were about the election.

The analysis of sentiment in the titles and text of posts leads to a more satisfying conclusion. The demeanor of fake news is likely to be negative. This conclusion is also in need of more data and exploration as the training data that was used was not specific to article titles or text. It was specific to human emotions felt in specific moments. To properly analyze fake news sentiment, a dataset that contains specific journalistic vocabulary and specific emotions related to journalism must be used.

This analysis could lead to a definite structure on analyzing fake news, as well a method to combat it by specifying the emotions that fake news articles are attempting to invoke. The internet is likely to become the main source of media, so this task will become especially important in the years to come.

Bibliography

Newman, Nic and Fletcher, Richard and Kalogeropoulos, Antonis and Levy, David A. L. and Nielsen, Rasmus Kleis, Reuters Institute Digital News Report 2017 (June 2017). Available at SSRN: https://ssrn.com/abstract=3026082

Wineburg, Sam and McGrew, Sarah and Breakstone, Joel and Ortega, Teresa. (2016). Evaluating Information: The Cornerstone of Civic Online Reasoning. Stanford Digital Repository. Available at: http://purl.stanford.edu/fv751yt5934

Dan-Glauser, E. S., & Scherer, K. R. (2013). The Difficulties in Emotion Regulation Scale (DERS): Factor Structure and Consistency of a French Translation. *Swiss Journal of Psychology, 72*(1), 5-11.

Yimin Chen , Niall J. Conroy , Victoria L. Rubin, Misleading Online Content: Recognizing Clickbait as "False News", Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection, November 13, 2015, Seattle, Washington, USA

Elisa Shearer, Jeffrey Gottfried, News Use Across Social Media Platforms 2017, Pew Research Center, September 7, September 7, 2017