Real News vs. Fake News:

Categorizing News Articles as Misinformation (Fake News)

And True (Real News)

Data 698

LeTicia Cancel

5/23/2023

Introduction

Information online is abundant through personal and professional blogs, local and global news websites, and free video services like YouTube. According to Siteefy.com, there are 197,046,670 active websites as of 9/18/2022 and "175 new websites created every minute"2. Articles from news websites like the New York Post published a variety of article types such as Celebrity Gossip, Entertainment, and Local News. There is an expectation from the reader that the information published is factual unless otherwise stated, like an opinion piece. How do we know the source is reliable if we read something from a site that is not as well known? Can we trust that the author thoroughly researched the topic prior to writing the article? Are all articles published on the web held to the same standard as a company such as the New York Times?

News Paper companies employ fact checkers who read draft articles and verify all information in the article before it is published. According to Mediabistro.com, “Fact-checkers help a source of news or information maintain credibility and integrity.” (<https://www.mediabistro.com/climb-the-ladder/skills-expertise/what-fact-checkers-do/>). Google has created its own electronic fact-checking system that allows the user to “search for stories and images that have already been debunked…)(<https://newsinitiative.withgoogle.com/resources/trainings/google-fact-check-tools/#lesson-section-5>). I would like to learn how online, automated fact-checkers work and what are some of the other methods used to identify misinformation work. What is the common thread between fake articles and how accurate is the algorithm used classify articles as fake? I understand that the most accurate way to determine if an article is fake is to run it through a fact-checking system or to have a professional editor check the author's sources for accuracy. However, most individuals do not have access to a fact-checking system and are not professional editors who will check the sources of an article they are reading. So, what can we do instead?

During Donald Trump's presidential campaign in 2016 and his tenure as the 45th president of the United States the term "fake news" became mainstream. According to The New Yorker "Judging from the President's tweets, his definition of 'fake news' is credible reporting that he doesn't like". The concept of fake news is not new or unique to Trump however it did become mainstream because it was used over one hundred and fifty times, as of December 3, 2017 [[citation](about:blank)]. The Cambridge Dictionary defines Fake News as "false stories that appear to be news, spread on the internet or using other media, usually created to influence political views or as a joke" [[citation](about:blank)]. We will use the Cambridge Dictionary definition to define Fake News and the term Real News to define the opposite, news that has been verified as truthful.

Social media has made it easier to spread information to a large group quickly. According to a study by Researchers at the Massachusetts Institute of Technology, tweets that contain false information are more likely to be retweeted and go viral than truthful information [[citation](about:blank)].

Facebook has been accused of creating an algorithm that prioritizes negative posts to a user's feed since people are likelier to interact with content that sparks a strong emotional reaction3. Using the pandemic as an example, it was a scary time for the world and the spread of misinformation about a new virus was dangerous and potentially deadly. Facebook updated its system to compare information against a fact-checker and flag posts as false4. Twitter has also attempted to stop the spread of misinformation by asking users to flag posts that "seem misleading"5. The existence of fake news is not new and not unique to the information about the pandemic that has been shared on social media in the past two years. What makes this so important today is how easily information is shared with many people. A system is needed to accurately identify misinformation as quickly as the information is spread and is needed across the web, not just on social media platforms. I am glad to see these social media companies attempting to identify and stop or slow the spread of misinformation by using fact-checkers and flagging by the community.

Literature Review

Fake news is a popular term, but do we really consume a large amount of fake news, or do we only consume a small amount, but it feels larger than it is because of the nature of the news? An article from Science.org [[citation](about:blank)] looked at the scale of misinformation in the media world. They first looked at which media types are used the most to consume information. They then examined how much of the information within the media type is misinformation. This information was also broken down by the age of the viewer. The study discovered that adults ages 18+ spent most of their day consuming non-news media on a television or mobile device. They spent an average of 20 minutes per day consuming media on Television. The information was broken up into age groups and the number of minutes per day steadily increased as the age groups increased. An interesting takeaway from this study was that although most information was consumed from what we would assume is a verifiable source, news outlets, fake news only comprised 1% of the overall news consumed.

An article from Stanford.edu [[citation](about:blank)] seeks to understand how misinformation is spread. Anecdotal information may make us point directly to social media, but this is not the only way news is consumed. The article mentions the game of telephone, which most of us played this game as children, and we still play it as adults, even though we might not think of it this way. When we consume information and feel compelled to share it, are we accurately communicating what we learned? If you read an article that upset you and shared this with a friend, how accurate would your explanation be? Researchers studied the spread of news through Twitter and found that when comparing the spread of a true and false news story, both reached 100 people, so this observation alone did not prove that fake news is spread more than real news. Instead, they found that fake news was "spread more easily because it was more infectious.”

Research Question

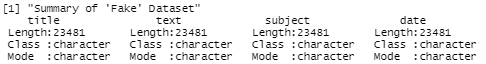
Fake news is not a new term, but it has become popular in the last 8 years. How can we easily identify whether what we read is real or fake news? If we can identify misleading news, what can we do about it? How can the publication of fake information be prevented? Is it possible to classify articles as Real News vs Fake News and how accurate is the classification without using a fact checker? For this project, I analyzed news datasets to identify fake vs. real news articles. People spend most of their time on the internet, so we are more likely to get our news from online articles instead of television. Information is spread quickly and easily through social media but how can we tell if the information we are reading is accurate? Is there a way to flag an article as misinformation? What are the consequences of an article being misrepresented as true? For this paper, I will use the term Fake News in reference to articles that are suspected to be misinformation and Real News in reference to articles with facts.

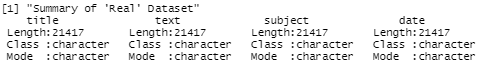
I plan to build a model to categorize the information as Real News or Fake News. The purpose of the model is not to check an article for factual accuracy but instead to flag an article as possible misinformation or Fake News. This flag can help the reader to make an informed decision about what they are reading. It will be used with public article datasets found on Kaggle that are assumed to be "Real News" to determine accuracy.

Data and Variables

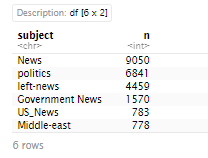
Data collection for this topic is challenging because you have to rely on someone else to identify and flag fake news for you and trust that their judgement is correct. The main data source for this project will come from Kaggle. The Fake.csv and True.csv file are datasets of news articles that have been identified as misinformation through fact-check research and a set of articles that have been verified as truthful. The articles are a bit old and the publication dates range from 2015 – 2018 so I searched for another dataset with more recent articles. The New York Post has a section on its website identifying fake news stories. The New York Post articles are technically not Fake News since it is not attempting to present misinformation but instead highlight stories that are known to be rumors. The articles are more recent and are from 2018 – 2023. This dataset can be used to understand better the type of topics used in fake news stories.

Each Kaggle file has an identical structure which will make the cleaning step simpler. The Fake dataset contains 23,481 observations and the Real dataset contains 21,417 observations. Each also contains four variables; Title and Text which are free text, Subject is categorical, and one date variable.

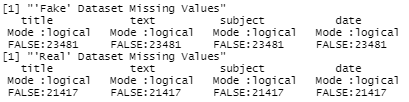




When we compare the count of Subjects for each dataset, the fake news dataset on the left has more categories with the "News" being the largest. The real news dataset on the right only has two categories with Political News being the largest.

If we check for missing data in both datasets, we see that there is no missing data.



The New York Post (NYP) dataset is a collection of articles scraped from the Fake News section of the New York Post website. Each page on the site has 20 articles so the screen scrape code went through 10 pages and collected a total of 200 articles. The Article Title, publish date, URL, and text were put into a dataframe. The NYP data has no subject column, but we can compare the top words from all three datasets. The top words for the Fake News dataset are Trump, President, People, White, Donald, and Image. The top words for the Real News dataset are Trump, U.S. Reuters, Republican, President, and House. The top words for the NYP dataset are Trump, People, News, Media, Fake, and Facebook. Since the Fake News and Real News datasets are articles from 2015-2017, the timeframe overlaps with Trump’s presidency, it is unsurprising to see his name and other names and terms related to that election in the charts. Two of the top words in the NYP dataset are fake and news which we will filter out and run again. The tags on the New York Post articles use the words fake and news so this does not give us any insight into this dataset. After removing the words fake and news, social and Twitter are now two of the top words for the New York Post.

A picture containing text, screenshot, diagram, colorfulness

Description automatically generated

All three datasets are then put into a Corpus and Text Mining functions are used to case fold (make all words lowercase) and remove stop words, punctuations, and symbols not needed for this analysis. The corpus is then put into a Document Term Matrix (DTM) to list all occurrences of words in the corpus (each word is put in its own column). After viewing the words in each DTM, additional filtering was needed to remove words not picked up from the TM function such as Post, go, can, $, and other symbols not needed for this analysis. The DTMs will be used again later in this analysis for the Topic Model.

Fake News DTM

A screenshot of a computer

Description automatically generated with medium confidence

Real News DTM

A screenshot of a computer

Description automatically generated with medium confidence

New York Post DTM

A screenshot of a computer

Description automatically generated with medium confidence

Sentiment Analysis

We will analyze sentiment to compare each dataset’s mood using the get\_nrc\_sentiment function. The sentiments used in this analysis are Anger, Anticipation, Disgust, Fear, Joy, Negative, Positive, Sadness, Surprise, and Trust. All three datasets share the top 3 moods, Positive, Negative, and Trust. All three barcharts look similar but the Fake News and NYP Sentiments are the closest matches.

A picture containing text, screenshot, colorfulness, diagram

Description automatically generated

Next, we compare the percentage of each emotion for the same datasets. The order of highest percentage to lowest for each emotion is the same for the Fake News dataset and the New York Post dataset. The Real News dataset differs from the second highest emotion down.

A picture containing text, screenshot, diagram, line

Description automatically generated

A picture containing text, screenshot, diagram, line

Description automatically generated

A picture containing text, screenshot, diagram, line

Description automatically generated

ADD MORE HERE ABOUT THE SENTIMENT ANALYSIS BEING TO POSITIVE

Statistical Methods

The Document Term Matrix (DTM) for each dataset cleaned further to remove terms that had a low frequency. The Latent Dirichlet Allocation (LDA) model was used to extract the top topics within each dataset. The top topics were graphed with the top terms for each group based on the Beta score. The Beta represents the density of the topic word. The higher the value, the more documents are composed of this word. Although we do not know exactly what was written in each article for all datasets, the LDA topic models help give us a better understanding of what the top articles were about.

Most groups in the Fake News dataset have terms related to the government and election with high beta scores. We know that the publish dates for the articles in this dataset were during a presidential election, so it makes sense that we see many terms related to President Trump’s election. We also see that topic #9 includes September, Hurricane, and Puerto Rico. Two major hurricanes struck Puerto Rico in September 2017 and while the date of that event lines up with this dataset, it appears as part of the Fake News dataset.

A screenshot of a graph

Description automatically generated with low confidence

The top topics and terms in the LDA model for the Real News dataset is like the LDA for the Fake News dataset. Most of the topics are related to the election and the government. The topics with terms that have a higher Beta score are topics 2 and 5 so the terms in both topics have a high density. Many of the terms in both topics overlap – Trump, U.S., Russia, President, and Said, can be found in both topics. Topic 12 has lower density scores and some of the terms found in this group are Tax, Percent, Rate, and Reform.

A screenshot of a graph

Description automatically generated with low confidence

The LDA model for the New York Post dataset does share some similarities with the previous LDA models, but there are new topics in this set. Topic 10 has the terms Queen, Elizabeth, Death, and Tweet, which informs us that one of the topics of more recent fake news was related to the passing of Queen Elizabeth. Topic 6 appears to be related to the American Commentator Joe Rogan, and the first topic appears to be related to college students.

A screenshot of a graph

Description automatically generated with low confidence

The LDA models provide insight into the main topics for all three datasets, but is it possible to build a model that can flag an article as fake news? The fake and real datasets were used to build the next model since they are both categorized. The fake dataset contains 23,481 articles and the real dataset contains 21,417 articles. If we combine both datasets, we get a total of 44,898 articles. This is a great amount of data, however, both datasets had to be reduced the 5,000 articles each. Anything over 10,000 articles negatively impacted the performance of R Studio and caused it to freeze.

Results

The DTM for each dataset was split into a 70/30 train and test set and three models were created and tested to see if we can classify an article as Fake or Real. The models built for each dataset were Naives Bayes, Random Tree, and Support Vector Machine (SVM). The results for the Naives Bayes model show close to a 50/50 split in its prediction which does look like an accurate prediction.

A screenshot of a computer

Description automatically generated with low confidence

The Cross Table for the Random Forest model has similar results to the Naives Bayes model prediction. It is close to a 50/50 split.

A screenshot of a computer

Description automatically generated with low confidence

Lastly, we have the Cross Table for the SVM model which is also similar to the prior two models.

A screenshot of a computer

Description automatically generated with low confidence

Conclusion

When searching for public datasets on news articles, most datasets were a collection of articles published 5 years ago or later. 2017 was the start of Donald Trump's presidency so the biggest news articles are more likely to be related to the result of the election. This project could be improved if more recent articles were used. I would put together my own dataset of articles published within the last 2 years by creating a function in R to screen scrape article text from various news websites. It would be a mix of articles from websites like The New York Times and articles from satirical websites The Onion.

R Studio can’t handle 10,000 rows for this type of analysis. Need a machine with more power and memory to handle the demand of the program.

References

1. [https://www.newyorker.com/magazine/2017/12/11/donald-trumps-fake-news-tactics](about:blank)
2. [https://dictionary.cambridge.org/us/spellcheck/english/?q=real+news](about:blank)
3. [https://www.science.org/content/article/fake-news-spreads-faster-true-news-twitter-thanks-people-not-bots](about:blank)
4. [https://www.science.org/doi/10.1126/sciadv.aay3539](about:blank)
5. [https://news.stanford.edu/2021/10/25/foil-fake-news-focus-infectiousness/](about:blank)

GENERAL COMMENTS

LeTicia,

You’ve made progress, but still have some work to do on your draft.

Given the (late) timeframe, I’ve approved your draft and posted it in the “Final-Approved Mid-Term Drafts” folder in the “Feedback on Drafts” section on the course site – with the understanding that you continue to work on the clarity and precision of the language, which I hope to see addressed in your final draft.

The paragraph that starts on the bottom of the first page does a good job explaining your research topic, which I would use as your intro/opening (i.e. this is an exploratory study to see if there might be consistent criteria or a fixed set of defining features to differentiate “fake” from “real” news, given new/emerging news datasets that have been trained to flag news articles as potentially false or true).

Citations need some work. As per the AMA citation style that you’ve chosen, you don’t need to embed links in the body text, just a number in numerical order; number of the citation is put in brackets, not in superscript, as would a footnote; and the citations appear to be mis-numbered.

Hope this helps. Please let me know if you want to discuss.

A