Real News vs. Fake News:

Categorizing News Articles as Misinformation (Fake News) and True (Real News)

Data 698

LeTicia Cancel

05/21/2023

Introduction

Information online is abundant and easily accessible from 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"3. 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 lesser-known? Can we trust that the author did their research 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?

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." Fake news is not new or unique to Trump; however, it became mainstream because it was used over one hundred and fifty times as of December 3, 2017. 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"2. 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 information4. 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 reaction5. Using the pandemic as an example, this was such a scary time for all, and spreading 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 false5. Twitter has also attempted to stop the spread of misinformation by asking users to flag posts that "seem misleading"6. Fake news is neither new nor unique to the pandemic information shared on social media in the past 2-3 years. What makes this topic so important today is how easily information is shared with many people. A system is needed to accurately identify misinformation as quickly as this information is spread and is needed across the web, not just on social media platforms.

It is good 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. Learning how the fact-checker and other methods of identifying misinformation work would be beneficial. What is the common thread between these fake news articles, and how accurate is the algorithm used to catch the misinformation? 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. Most individuals need access to a fact-checking system and are not professional editors who would check the sources of an article we are reading. So, what can we do instead?

Literature Review

Fake news is such a popular term but do we really consume that much fake news or do we only consume a small percentage and it feels great because it is a big deal? An article from Science.org examined the scale of misinformation in the media world7. 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. Their study found that adults aged 18+ spent most of their day consuming non-news media, which was consumed on television or mobile devices. The average number of minutes per day spent consuming Television news was 20. They broke it 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 seeks to understand how misinformation is spread8. Anecdotal information may make us point directly to social media, but this is not the only way news is consumed. They mention the game of telephone, which we all played as a child, and we still play it as adults, even though we might not think of it in 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 topic but 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 it be prevented? Can we classify articles as Real News or Fake News and how accurate can it be without a fact checker? For this project, I chose to analyze news datasets to identify true versus fake information, or as it is sometimes described on social media, Real News vs. Fake News. 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 or are suspected to be misinformation and Real News in reference to articles with factual information.

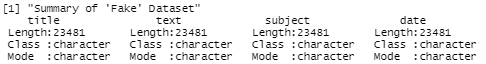
I plan to build a model to categorize the information as Real News or Fake News. The purpose of the model is not to check an article for factual accuracy but instead to flag an article as possible misinformation or Fake News. This flag can help the reader make an informed decision about what they are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

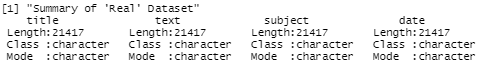
Data and Variables

The data source for this project will come from Kaggle and the New York Post. The Kaggle datasets are made up of two files, Fake.csv and True.csv. The files are a collection of news articles identified as misinformation (Fake) through fact-checking research and a set of articles verified as truthful (True). The articles were published from 2015 -2018 so we need a third dataset with more recent articles. Since a collection of more recent articles is unavailable, we will scrape articles from the New York Post Fake News website.

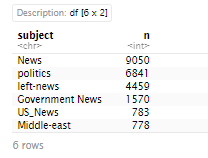
200 articles were scraped from the New York Post website. The article data includes the article date, title, URL, and text, which was added to a dataframe. The Kaggle Fake dataset contains 23,481 observations and the Real dataset contains 21,417 observations. Each contains four variables; Title and Text which are free text, Subject is categorical, and one date variable. All three files have similar structures which will make the cleaning step simpler.

A picture containing text, font, line, screenshot

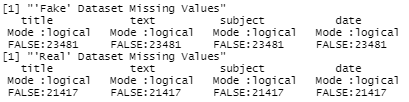
Description automatically generated



The fake and real datasets have a Subject column so we can compare the top subjects of each dataset. The fake news dataset on the left has more subjects 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 are no missing data so we do not have to do further cleaning for this.



The New York Post dataset 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, Obama, Donald, and Clinton. The top words for the Real News dataset are Trump, U.S. Reuters, President, House, and Government. Since both datasets are articles from 2015-2017 that overlap with the Trump presidency, it is unsurprising to see his name and other names and terms related to that election. Two of the top words in the New York Post dataset are the words 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 tell give us any insight to this dataset.

A picture containing text, screenshot, diagram, colorfulness

Description automatically generated

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.

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

We will do a sentiment analysis 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. The

A picture containing text, screenshot, colorfulness, diagram

Description automatically generated

Statistical Methods

The Corpus package and the Text Mining Package is used put the text together and clean the text. The words will be rated using Tf-idf to find the most important words. Two models will be created and the results compared. Naïve Bayes model and Random Forest. The most accurate model will be used with real news articles.

Results

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.

References

1. Coll, S. (2017, December 3). Donald Trump's "Fake news" tactics. The New Yorker. Retrieved December 11, 2022, from <https://www.newyorker.com/magazine/2017/12/11/donald-trumps-fake-news-tactics>
2. Real+News - did you spell it correctly. alternative spellings in the British english dictionary - cambridge dictionary (US). real+news - Did you spell it correctly. Alternative spellings in the British English Dictionary - Cambridge Dictionary (US). (n.d.). Retrieved December 11, 2022, from <https://dictionary.cambridge.org/us/spellcheck/english/?q=real%2Bnews>
3. Huss, N. (2022, August 22). How many websites are there in the world? (2022). Siteefy. Retrieved September 18, 2022, from https://siteefy.com/how-many-websites-are-there/
4. Fake news spreads faster than true news on Twitter-thanks to people, not bots. Science. (n.d.). Retrieved December 11, 2022, from <https://www.science.org/content/article/fake-news-spreads-faster-true-news-twitter-thanks-people-not-bots>
5. Mukul, P. (2021, May 28). Explained: How new facebook feature flags misinformation. The Indian Express. Retrieved September 10, 2022, from <https://indianexpress.com/article/explained/facebook-misinformation-fake-news-tool-7332659/>
6. Spangler, T. (2021, August 17). Twitter is asking users to flag misinformation, including about COVID and elections. Variety. Retrieved September 10, 2022, from <https://variety.com/2021/digital/news/twitter-users-flag-misinformation-covid-elections-1235043215/>
7. Evaluating the fake news problem at the scale of the ... - science. Evaluating the fake news problem at the scale of the information ecosystem. (n.d.). Retrieved December 12, 2022, from <https://www.science.org/doi/10.1126/sciadv.aay3539>
8. University, S. (2022, April 11). To foil fake news, focus on infectiousness. Stanford News. Retrieved December 11, 2022, from <https://news.stanford.edu/2021/10/25/foil-fake-news-focus-infectiousness/>