Real News vs Fake News:

Categorizing News Articles as Misinformation (Fake News)

And True (Real News)

Data 698 Midterm Project

LeTicia Cancel

10/29/22

Introduction

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 term Real News to define the opposite, news that has been verified as truthful.

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. If the information we are reading is presented as factual, how do we know that the source is reliable? Can we read something and trust that the author did their research prior to writing the article? Are all articles published on the web held the same standard as a company such as the New York Times?

Social media has made it easier to spread information quickly to a large group of people. According to a study conducted by the 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 more likely to interact with content that sparks a strong emotional reaction3. Using the pandemic as an example, this was such a scary time for all and the spread of misinformation about a new virus was dangerous and potentially deadly. Facebook updated their 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 is also not unique to the pandemic information that has been shared on social media in the past two years. What makes this so important today is just how easily information is shared to a large group of 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.

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. I would like to learn how the fact-checker and other methods of identifying misinformation work. What is the common thread between these "fake" 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, including myself, do not have 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 [[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 looked out how much of the information within the media type is misinformation. This information was also broken down by age of viewer. Their study found that adults ages 18+ spent most of their day consuming non-news media and this non-news media was consumed on television or mobile device. The average number of minutes per day spent consuming Television news was 20 minutes. 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 made up 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 his 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 this game as adults, even though we might not think of it in this way. When we consume any information and feel compelled to share it, are we communicating what we learned accurately? 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 what they found was that fake news was "spread more easily because it was more infectious".

Research Question

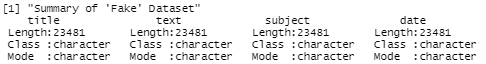
Fake news is not a new topic but it has become a popular term in the last 8 years. How can we easily identify if what we are reading is real or fake news? If we can identify misleading news, what can we do about it? Can it be stopped? Can we classify articles as Real News vs Fake News and how accurate can it be without the use of 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.

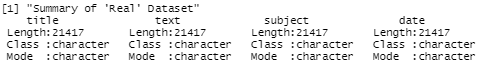
I plan to build a model that can categorize the information as Real News or Fake News. The purpose of the model is not to check an article for factual accuracy but instead flag an article as possible misinformation or Fake News. This flag can help the reader to make an informed decision with what they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

Data and Variables

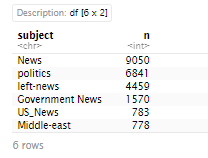
Data collection for this topic is challenging since you must rely on someone else to identify and flag fake news for you and trust that their judgement is correct. The main data source for this project will come from Kaggle. The Fake.csv and True.csv file are datasets of news articles that have been 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 range from 2015 – 2018 so I searched for a second dataset with more recent articles. The second dataset comes from the New York Post where they focus on rumors instead of attempting to create an article and pass it off as Real News. Since the New York Post articles are technically not Fake News and are not attempting to present misinformation. Instead, it highlights stories that they know to be rumors. The articles are more recent and are from 2018 – 2023.

Both Kaggle files have identical structures which will make the cleaning step simpler. The Fake dataset contains 23,481 observations and the Real dataset contains 21,417 observations. Each also contain 4 variables; Title and Text which are free text, Subject is categorical, and one date variable.

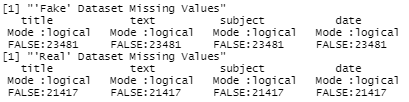




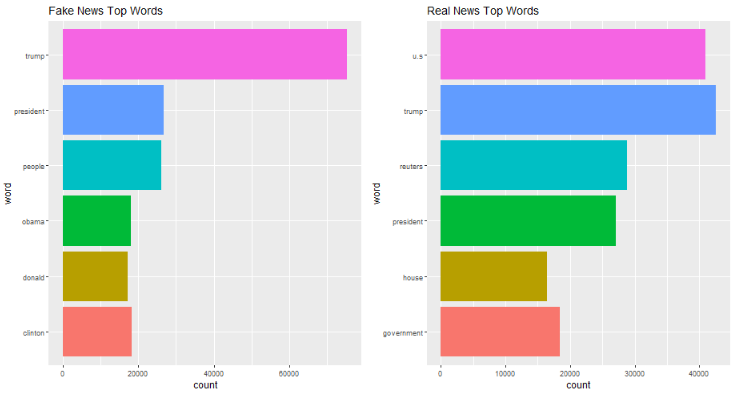
When we compare the count of Subjects for each dataset, the fake news dataset on the left has more categories with the "News" being the largest. The real news dataset on the right only has two categories with Political News being the largest.

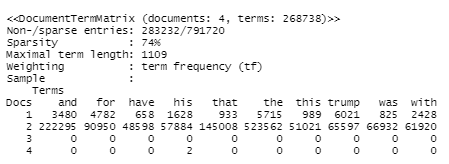
If we check for missing data in both datasets we see that there are no missing data so we do not have to do further cleaning for this.



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 which overlaps with the Trump presidency, it is not surprising to see Trump's name and other names and terms related to the past election.

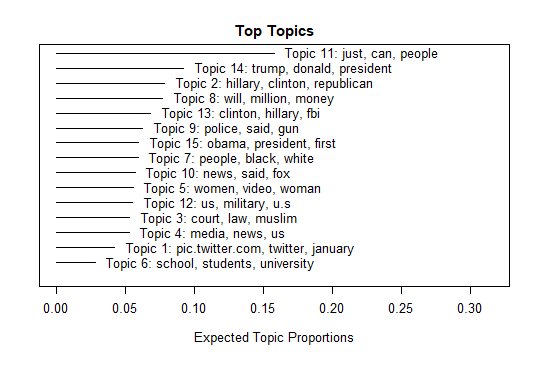


Each dataset is then put into a Corpus and the Text Mining functions are used to case fold (make all words lowercase), and stemming, the remove the end letters of words and keep the main word such as "Sleeping" and translated to "Sleep". The corpus is then put into a Document Term Matrix (DTM) to list all occurrences of words in the corpus (each word is put in its own column). The removeSparseTerms function is used and set to 0.99 to remove more words that are not significant to this analysis.

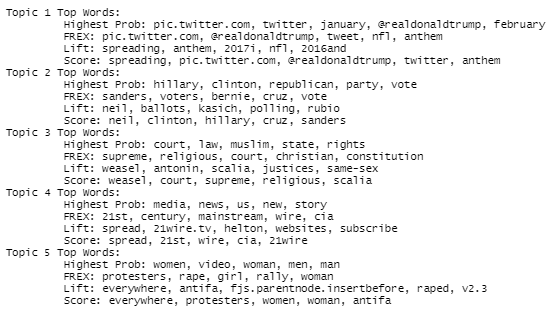


Now that we know the top words in this dataset we can identify important topics. Document Feature Matrix (DFM) is another useful tool in exploring this dataset. After creating the corpus and cleaning the data with DFM we can compare the top words and see if we get similar results. Trump, President, and People are top words from the fake dataset. 

Using Structural Topic Modeling we can plot the top topics from the datasets. The top topics contains words that are not very significant and a couple of them can probably be removed. The second top topic is not is not surprising as the words are also the most frequently occuring words in the dataset.



We will now focus on just the top 5 topics and their FREX weighted top terms. The terms listed in the FREX are terms that are common for that topic and/or terms that are exclusive for that topic. Topic 2 is the easiest to interpret with the FREX. Topic 2 is 'hillary', 'clinton', and 'republican', names and terms that are related to government and an election. The FREX for this group is also related to the same subject.



Now we can focus just on the FREX for the top 20 features and calculate the conditional probability that a topic is prevalent in a given document. Topic 1 is the most prevalent in the first document with a probability of 0.325.







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