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"[1]. 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 publication. According to Mediabistro.com, “Fact-checkers help a source of news or information maintain credibility and integrity.”[2] Google has created its own electronic fact-checking system that allows the user to “search for stories and images that have already been debunked…).”[3] 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 to 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.[4] 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"[5]. 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[6]. 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 reaction[7]. 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. Twitter is attempting to stop the spread of misinformation by asking users to flag posts that "seem misleading"[8]. 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 examined 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[9].

An article from Stanford.edu 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."[10]

Research Question

Fake news is not new, 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 likelier to get our news from online articles than 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 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 make an informed decision about

what they are reading. We will use 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 judgment is correct. The main data source for this project will come from Kaggle. The Fake.csv and True.csv files 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 so it is still useful for this project. 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.

```
[1] "Summary of 'Fake' Dataset"
    title              text               subject             date
 Length:23481       Length:23481       Length:23481       Length:23481
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
[1] "Summary of 'Real' Dataset"
    title              text              subject             date
 Length:21417       Length:21417       Length:21417       Length:21417
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

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.

Description: df [6 x 2]

subject <chr>	n <int>
News	9050
politics	6841
left-news	4459
Government News	1570
US_News	783
Middle-east	778

6 rows

Description: df [2 x 2]

subject <chr>	n <int>
politicsNews	11272
worldnews	10145

2 rows

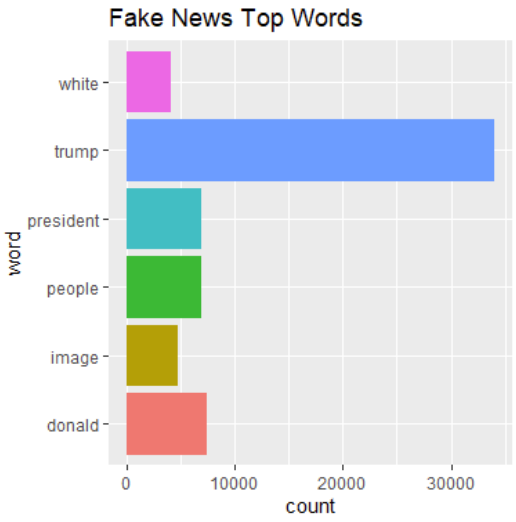
If we check for missing data in both datasets, we see that there is no missing data.

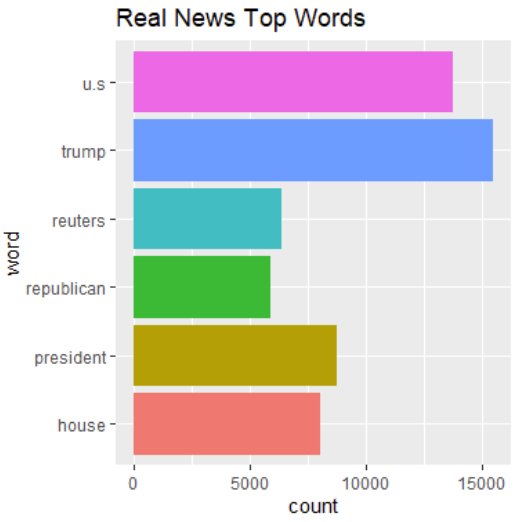
```
[1] "'Fake' Dataset Missing Values"
   title              text               subject            date
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:23481     FALSE:23481     FALSE:23481     FALSE:23481
[1] "'Real' Dataset Missing Values"
   title              text               subject            date
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:21417     FALSE:21417     FALSE:21417     FALSE:21417
```

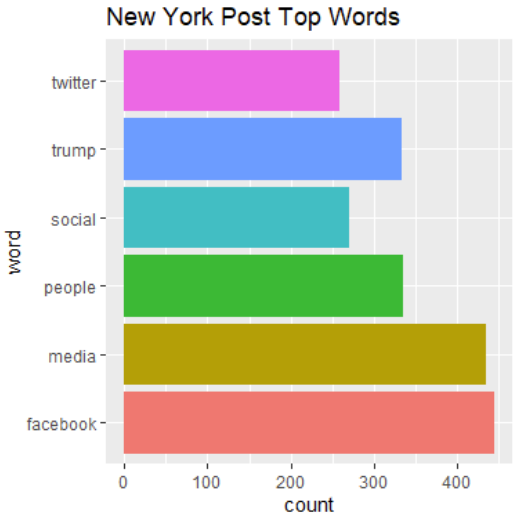
The New York Post (NYP) dataset is a collection of articles scraped from the Fake News section of the New York Post website. Each page on the site has 20 articles so the screen scrape code went through 10 pages and collected a total of 200 articles. The Article Title, publish date, URL, and text were 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, and it is unsurprising to see names and terms related to that election in the charts. Two of the top words in the NYP dataset were fake and news so we filtered out these terms and ran the chart again. The tags on the New York Post articles use the words fake and news so this only inflates the terms and 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.







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 topic modeling.

Fake News DTM

Description: df [6 × 2]

	word <chr>	freq <dbl>
trump	trump	34365
imag	imag	7813
donald	donald	7516
presid	presid	7367
peopl	peopl	6898
said	said	6358

6 rows

Real News DTM

Description: df [6 × 2]

	word <chr>	freq <dbl>
said	said	23022
trump	trump	15260
presid	presid	9004
republican	republican	8709
hous	hous	8119
state	state	8028

6 rows

New York Post DTM

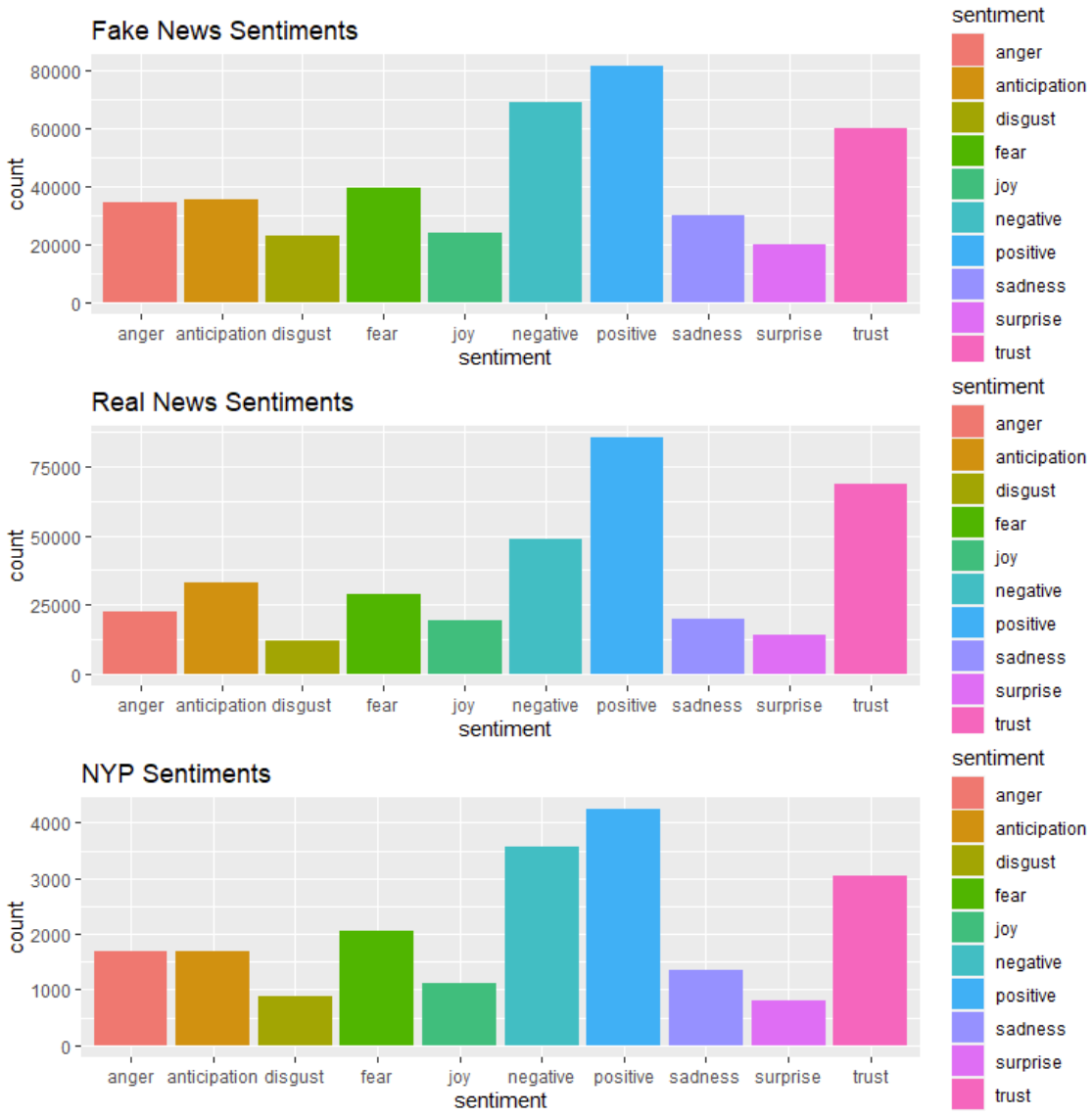
Description: df [6 × 2]

	word <chr>	freq <dbl>
media	media	383
facebook	facebook	315
peopl	peopl	291
report	report	278
’s	’s	264
say	say	255

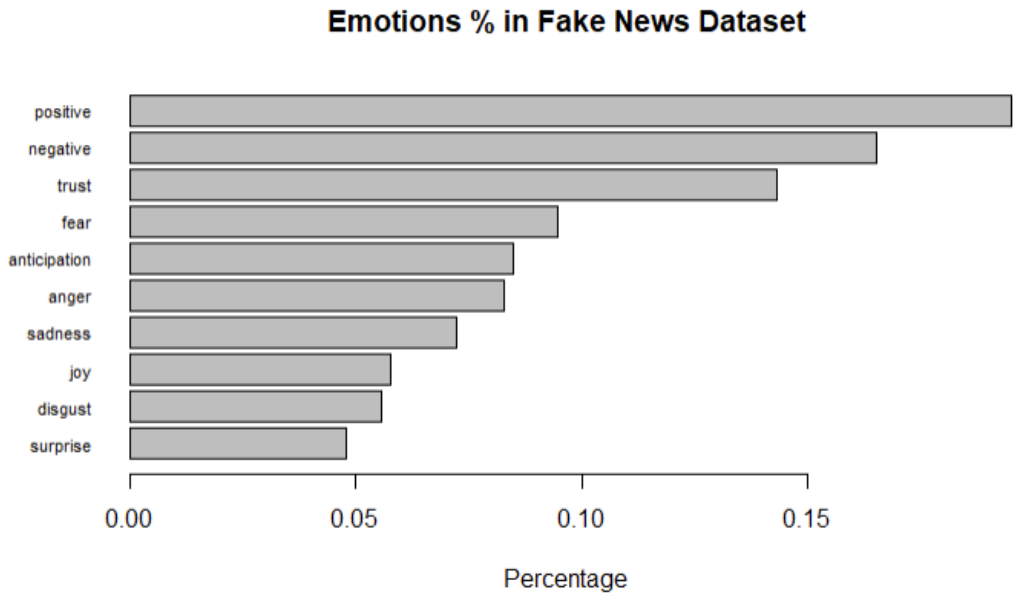
6 rows

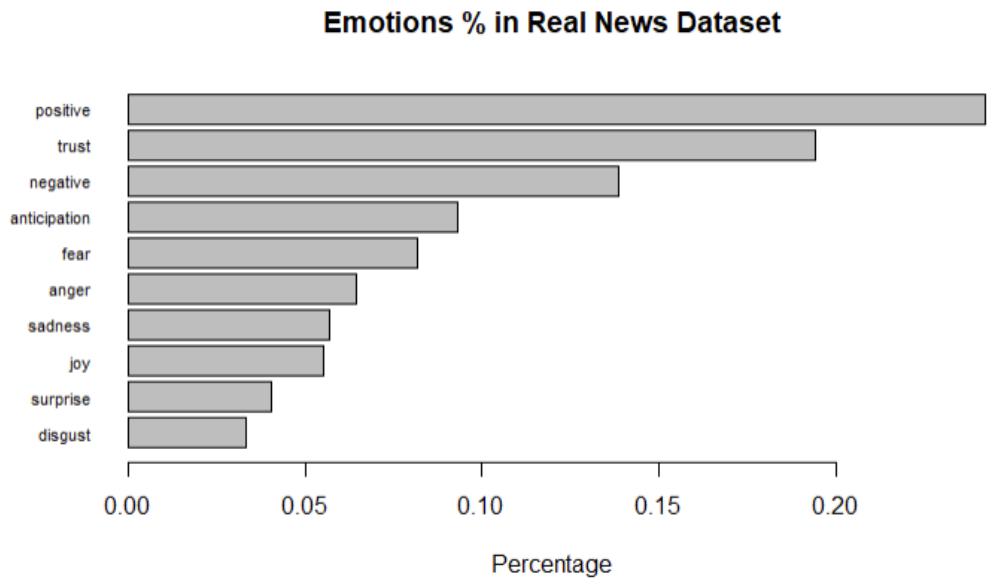
Sentiment Analysis

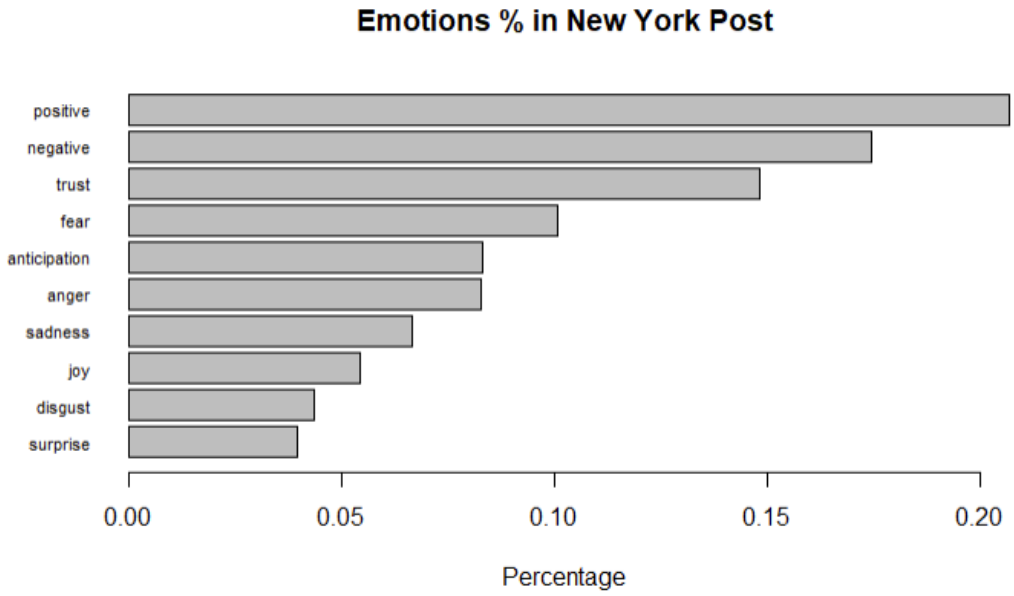
We will analyze sentiment to compare each dataset’s mood using the get_nrc_sentiment function. The sentiments used in this analysis are Anger, Anticipation, Disgust, Fear, Joy, Negative, Positive, Sadness, Surprise, and Trust. All three datasets share the top 3 moods, Positive, Negative, and Trust. The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


Fake News Sentiments
count
80000
60000
40000
20000
0
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust
sentiment
Real News Sentiments
count
75000
50000
25000
0
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust
sentiment
NYP Sentiments
count
4000
3000
2000
1000
0
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust
sentiment
sentiment
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust
sentiment
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust
sentiment
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust

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. It is interesting to see that the top emotion in all three datasets is Positive and Trust is either second or third depending on which chart you are reviewing. Most people would not associate a fake news article with a Positive emotion. This comparison helps highlight the difficulty a reader might face in identifying a fake news article based solely on the tone.



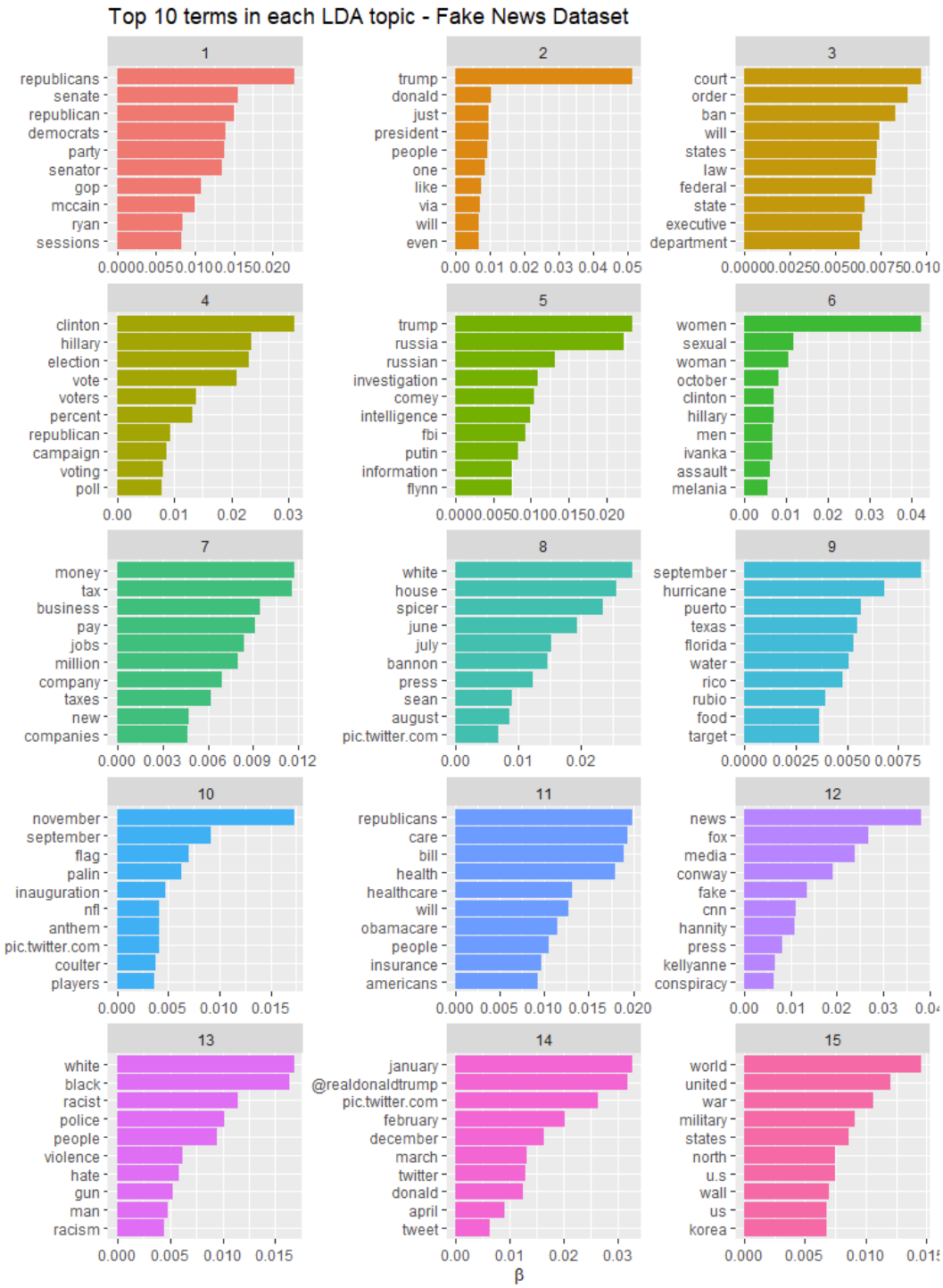

Emotions % in Real News Dataset
positive
trust
negative
anticipation
fear
anger
sadness
joy
surprise
disgust
0.00
0.05
0.10
0.15
0.20
Percentage


Emotions % in New York Post
positive
negative
trust
fear
anticipation
anger
sadness
joy
disgust
surprise
0.00
0.05
0.10
0.15
0.20
Percentage

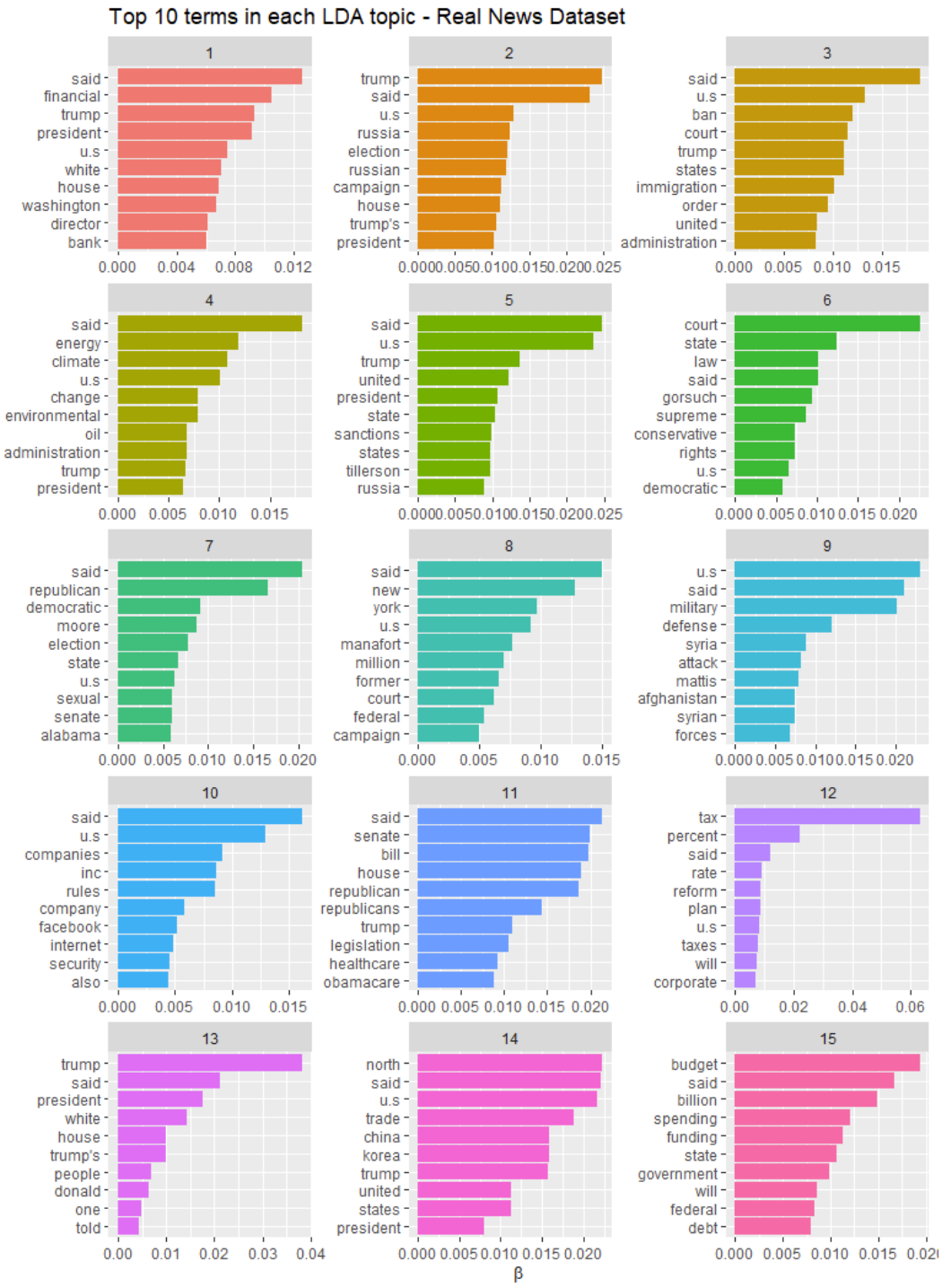
Statistical Methods

The Document Term Matrix (DTM) for each dataset was 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.

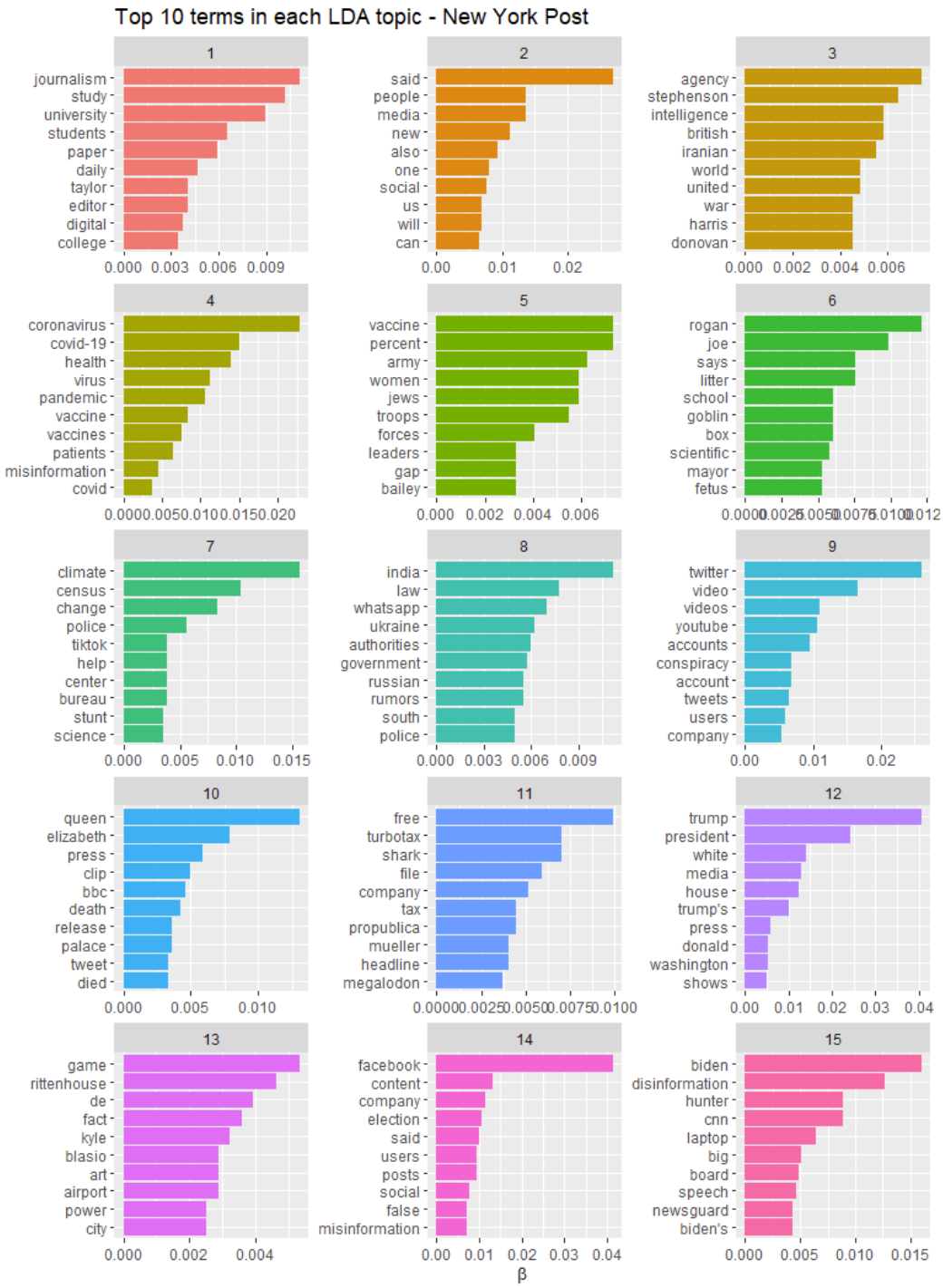

Top 10 terms in each LDA topic - Fake News Dataset
1
republicans
senate
republican
democrats
party
senator
gop
mccain
ryan
sessions
0.000 0.005 0.010 0.015 0.020
2
trump
donald
just
president
people
one
like
via
will
even
0.00 0.01 0.02 0.03 0.04 0.05
3
court
order
ban
will
states
law
federal
state
executive
department
0.00000 0.00250 0.00500 0.00750 0.010
4
clinton
hillary
election
vote
voters
percent
republican
campaign
voting
poll
0.00 0.01 0.02 0.03
5
trump
russia
russian
investigation
comey
intelligence
fbi
putin
information
flynn
0.0000 0.0050 0.0100 0.0150 0.020
6
women
sexual
woman
october
clinton
hillary
men
ivanka
assault
melania
0.00 0.01 0.02 0.03 0.04
7
money
tax
business
pay
jobs
million
company
taxes
new
companies
0.000 0.003 0.006 0.009 0.012
8
white
house
spicer
june
july
bannon
press
sean
august
pic.twitter.com
0.00 0.01 0.02
9
september
hurricane
puerto
texas
florida
water
rico
rubio
food
target
0.0000 0.0025 0.0050 0.0075
10
november
september
flag
palin
inauguration
nfl
anthem
pic.twitter.com
coulter
players
0.000 0.005 0.010 0.015
11
republicans
care
bill
health
healthcare
will
obamacare
people
insurance
americans
0.000 0.005 0.010 0.015 0.020
12
news
fox
media
conway
fake
cnn
hannity
press
kellyanne
conspiracy
0.00 0.01 0.02 0.03
13
white
black
racist
police
people
violence
hate
gun
man
racism
0.000 0.005 0.010 0.015
14
january
@realdonaldtrump
pic.twitter.com
february
december
march
twitter
donald
april
tweet
0.00 0.01 0.02 0.03
15
world
united
war
military
states
north
u.s
wall
us
korea
0.000 0.005 0.010
β

The top topics and terms in the LDA model for the Real News dataset are 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 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.



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.



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 were 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 classification models built for each dataset were Naive Bayes, Random Tree, and Support Vector Machine (SVM). The results for the Naive Bayes model show close to a 50/50 split in its prediction which does not look like an accurate prediction.

Cell Contents
N
N / Row Total
N / Col Total

Total Observations in Table: 3000

Prediction	Actual Fake	Real	Row Total
Fake	730	759	1489
	0.490	0.510	0.496
	0.487	0.506	
Real	770	741	1511
	0.510	0.490	0.504
	0.513	0.494	
Column Total	1500	1500	3000
	0.500	0.500	

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.

Cell Contents
N
N / Row Total
N / Col Total

Total Observations in Table: 3000

Prediction	Actual Fake	Real	Row Total
Fake	725	757	1482
	0.489	0.511	0.494
	0.483	0.505	
Real	775	743	1518
	0.511	0.489	0.506
	0.517	0.495	
Column Total	1500	1500	3000
	0.500	0.500	

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

Cell Contents
N
N / Row Total
N / Col Total

Total Observations in Table: 3000

Prediction	Actual: Fake	Actual: Real	Row Total
Fake	725	758	1483
	0.489	0.511	0.494
	0.483	0.505	
Real	775	742	1517
	0.511	0.489	0.506
	0.517	0.495	
Column Total	1500	1500	3000
	0.500	0.500	

Conclusion

When searching for public datasets on news articles, it was difficult to find something with recent articles. The datasets were a collection of articles published at least 5 years ago and overlapped with the start of Donald Trump's presidency. Most topics would be related to the election and the elected president during this timeframe. Screen scraping to collect articles from websites is possible in R, but the challenge is finding websites that show articles categorized as fake.

The sentiment analysis showed more similarities than expected. Most would not associate a fake new article with the word Positive or Trust, but they were the highest and third highest emotion

respectively. The similarities in sentiment for all three datasets highlight how difficult it is for an individual to identify if the article they are reading is fake without the help of a fact checker.

We compared the topics of all three datasets using an LDA model and found that the groups and terms were also similar with some slight variation. Most topics in all three datasets had terms related to the government. The fake news dataset had some topics related to the hurricane in Puerto Rice and the NFL. The real news dataset had topics related to energy and the environment. The New York Post dataset had topics related to social media and conspiracies.

As with sentiment analysis, the LDA models also highlight the difficulty in identifying misinformation. If a fake news article and a real news article are both written in a positive tone and are about the same topic, how can a reader spot the fake if they are not already an expert in the topic they are reading about? The performance of the three models also shows the difficulty in accurately predicting of an article is real or fake. The mood and topics of all three datasets were so similar that more testing is needed to get an accurate prediction.

It is important to have a reliable fact-checking system for the Web to help reduce the spread of misinformation. This project could be improved if there were a repository of fake news articles, or a website devoted to sharing these articles for educational purposes. This project could also be improved by expanding the categories to include categories like satire which can be found on websites like The Onion. How would the model perform if we included articles labeled as satire, articles meant to be comedic and fake but not meant to deceive the reader?

References

1. 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/
2. Bey, Amirah. What Fact Checkers Do and Why the Role Is so Important. Oct. 2018, www.mediabistro.com/climb-the-ladder/skills-expertise/what-fact-checkers-do.
3. Google Fact Check Tools - Google News Initiative. newsinitiative.withgoogle.com/resources/trainings/google-fact-check-tools/#lesson-section-5.
4. 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
5. 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
6. 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
7. 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/
8. 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/

9. 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
10. 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/