Data Appendix

For this project, we took the two initial datasets we obtained from Kaggle, combined them, cleaned them, to create a single dataset for analysis called filtered_news.csv. The appendix serves to provide more information on this dataset.

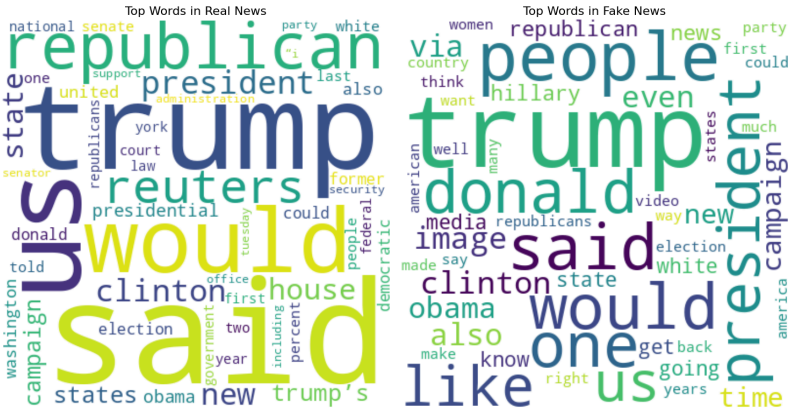
I. filtered_news.csv:

This dataset consists of 8 columns: [*"title", "text", "subject", "date", "True", "clean_text", "sentiment", "words"*]. Each row of the dataset represents a news article published by an online source, along with its text, publish date, subject, sentiment score, and whether it is classified as real or fake news. The dataset contains 14,181 entries.

There were no missing observations for any of the variables, so for the sake of simplicity, the variable descriptions will not mention that.

Details of Each Variable:

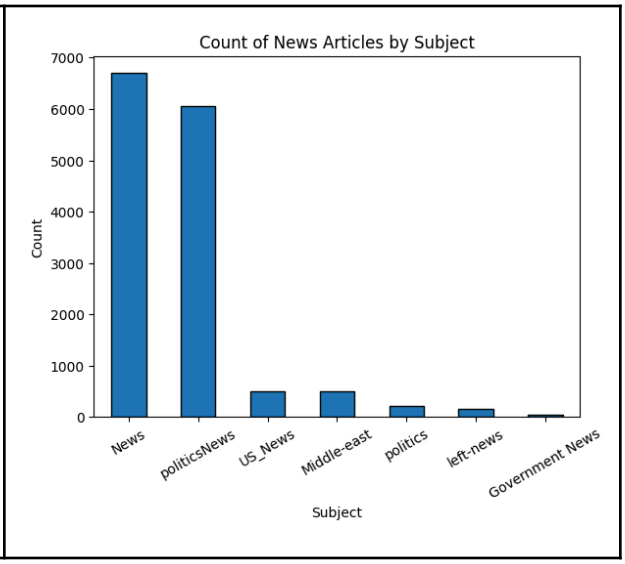
- title
  - Definition: the headline text of each article
  - Data type: text
  - Processing steps: None (kept initial data)
  - Response: Varies
- text:
  - Definition: the body text of each article
  - Data type: text
  - Processing steps: None (kept initial data)
  - Response: Varies
  - Word Cloud of Most Frequent Words:



- subject:
  - Definition: description of the category of the article
  - Data type: text, categorical
  - Processing steps: None (kept initial data)

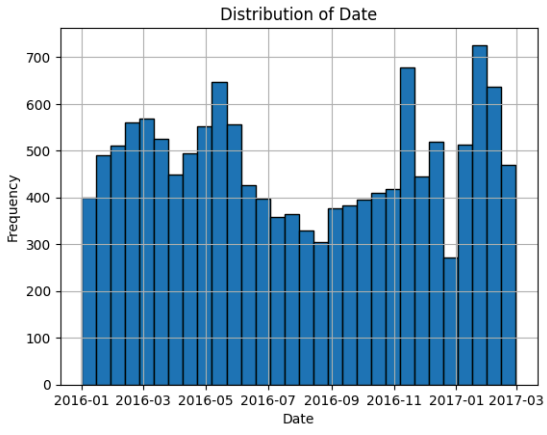
- Possible Responses: [World-News, Politics-News, Government-News, Middle-east, US News, left-news, politics, News]
- Frequency Information of Subject Areas

Subject	Count
News	6694
politicsNews	6057
US_News	502
Middle-east	497
politics	225
left-news	159
Government News	47



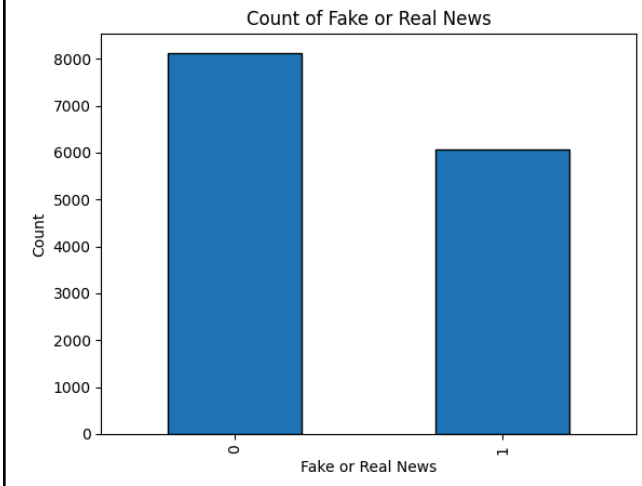
- date:
  - Definition: the publishing date of each article
  - Data Type: date object
  - Processing steps:
    a) In order to focus on the hypothesis, which focused on proximity to news near the 2016 election, the data was filtered to only include data entries between January 2016 to March 2017
    b) In order to perform this, we transformed the initial dates into a uniform date object.
    c) Then, we filtered the data entries to only include articles published between 01/01/16 and 03/01/2017
  - Response: Ranges from January 2016 to March 2017
  - Statistics:

	date
count	14181
mean	2016-07-30 23:48:25.436852224
min	2016-01-01 00:00:00
25%	2016-04-10 00:00:00
50%	2016-07-22 00:00:00
75%	2016-11-22 00:00:00
max	2017-02-28 00:00:00



- True:
  - Definition: indicates whether an article is fake or real
  - Data type: numerical, categorical
  - Processing steps:
    a) This column was added from the initial data to merge the fake and real news articles into a single dataset
    b) To do this, fake new articles were mapped to 0 and real new articles mapped to 1. From there, the two datasets were merged
  - Response: Values are either 0 or 1
    a) 0 signifies fake news, 1 signifies real news
  - Statistics:

	True
count	14181.000000
mean	0.427121
std	0.494678
min	0.000000
25%	0.000000
50%	0.000000
75%	1.000000
max	1.000000



- clean_text:
  - Definition: the processed text of each article after data cleaning function was applied
  - Processing steps:
    a) In order to create this column, a python function was applied to every entry, transforming the text from the *text* column to the *clean_text* column
    b) The python function used for clearing removed whitespace, URL formatting, and numbers in order to prepare for sentiment analysis
  - Response: Varies
- sentiment:
  - Definition: the sentiment score produced by the VADER python package for each article
  - Processing steps:
    a) To create this column, the VADER python package SentimentIntensityAnalyzer was applied to the clean_text column

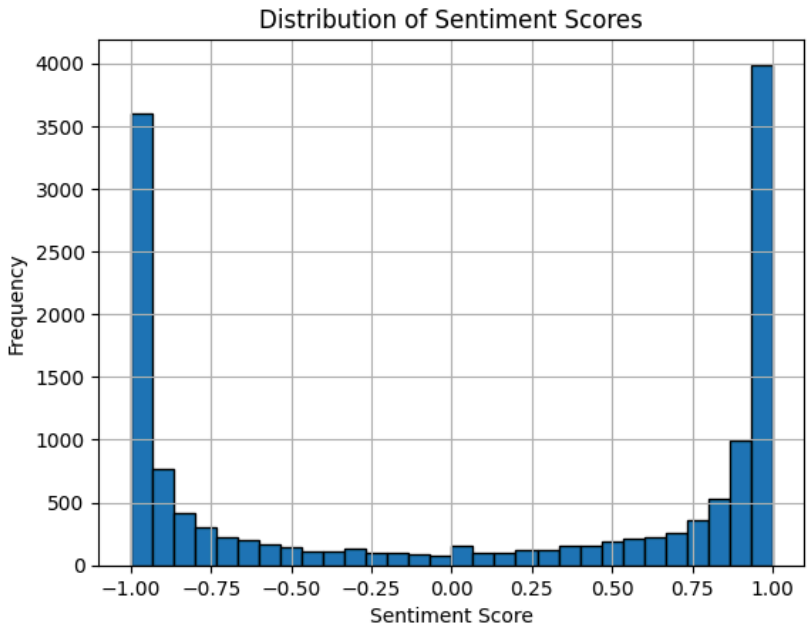
b) This produced a sentiment score ranging from -1 to 1 for every data entry

c)
```
# Initialize VADER Sentiment Analyzer
sia = SentimentIntensityAnalyzer()

# Apply sentiment analysis
filtered_df['sentiment'] = filtered_df['clean_text'].apply(lambda x: sia.polarity_scores(x)['compound'])
```

- Response: the sentiment scores ranged from -1 to 1, with scores closer to -1 reflecting more negative sentiment and scores closer to 1 reflecting more positive sentiment
- Statistics:

	sentiment
count	14181.000000
mean	0.058173
std	0.858291
min	-1.000000
25%	-0.937400
50%	0.308900
75%	0.950200
max	0.999800



- words:
  - Definition: the list of words from the clean_text column formatted as a list
  - Processing steps:
    a) In order to create this column, the text from clean_text columns were tokenized into separate words, formatted as a list
    b)
```
# Tokenize text into words
filtered_df['words'] = filtered_df['clean_text'].str.split()
```
  - Response: Varies