Predicting Fake and Real News

INSTRUCTOR: ADAM ANDERSEN

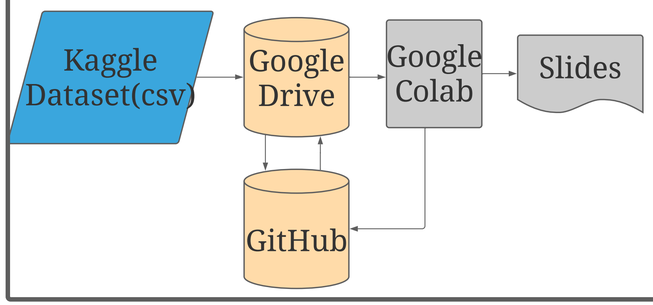
DIGITAL HUMANITIES, SUMMER 2021

STUDENT: JASPER ZHOU

Research Questions:

A) How has fake news changed?
B) Can we predict whether something is fake news or not using NLP?

Workflow:



Introduction:

Fake news is not a recent or unique issue, although it has been exacerbated by the ease and speed in which information could be spread. Since the 1890s, fake news has plagued the profession of journalism. Dubbed "yellow journalism", newspaper publishers would run sensationlistic news articles. However, fake news was eventually replaced by objective journalism, until recently due to the rise of the internet. (Source:UCSB)

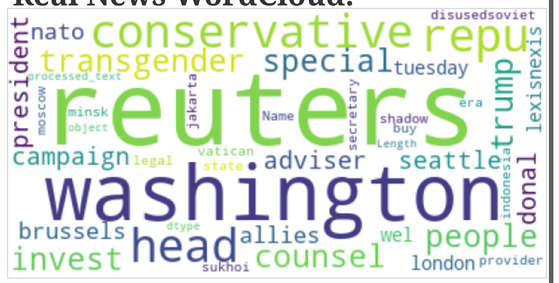
In terms of impacts fake news has on society, University of Michigan lists the following as possibilities:

- Anti-intellectualism
- Antiscience
- Widespread Mistrust

Works Cited:

- WordCloud Tutorial
- TextBlob Documentation
- Confusion Matrix tutorial
- Wikipedia for SVM
- Wikipedia for Logistic Regression

Real News WordCloud:



Fake News WordCloud:

EDA: *Continued on Page 2*

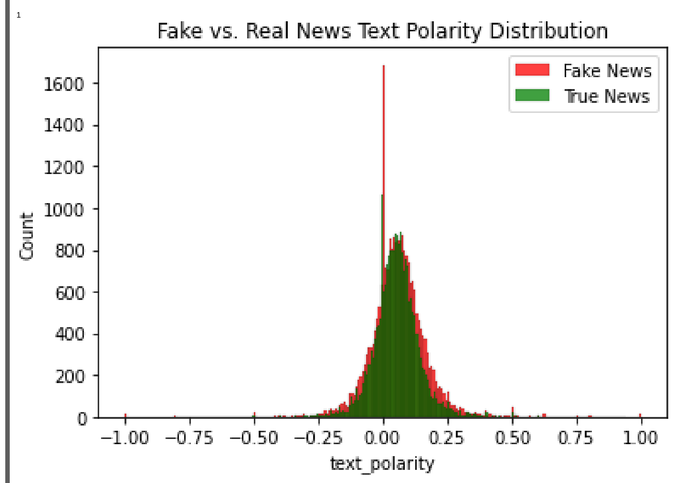
- From wordcloud, can see differences in words used
- Fake news dataset primarily about Donald Trump
- Sources matter!
  - Reuters makes up a large part of the real news data
  - 21st Century Wire, a right wing conservative web blog makes the a large part of fake news data

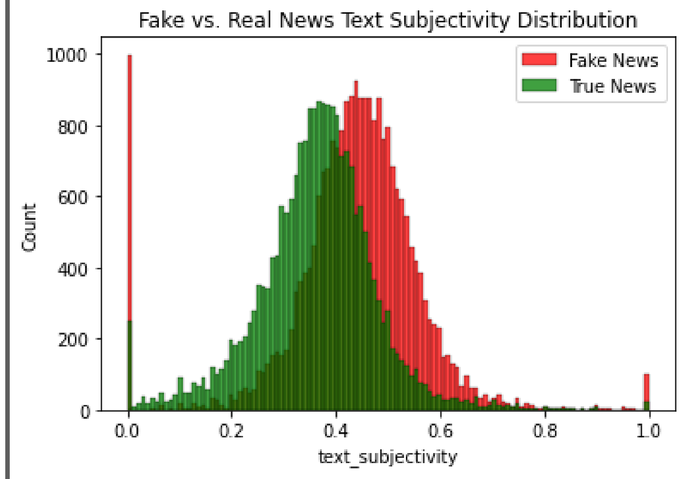
•Difference in topics discussed:
•Fake news tends to be nationalistic (america) and more focused on traditional issues (christmas, sheriff)
•Real news tends to be more about international issues (nato, brussels, london) and more on progressive ideals (transgender)

•Interestingly, it seems real news is more interested in conservatives than fake news is.

Sentiment Analysis: *Continued on Page 3*

•Tracked polarity and subjectivity scores
•Distribution of polarity scores were not markedly different in true and fake news
•Fake news contains more subjective vocabulary than true news

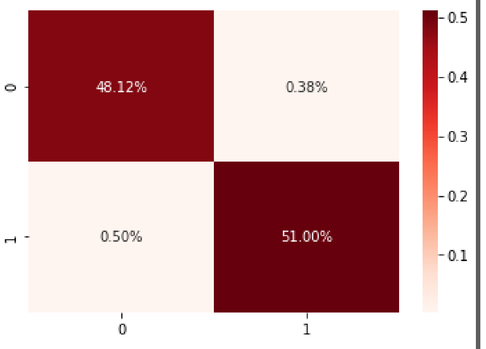




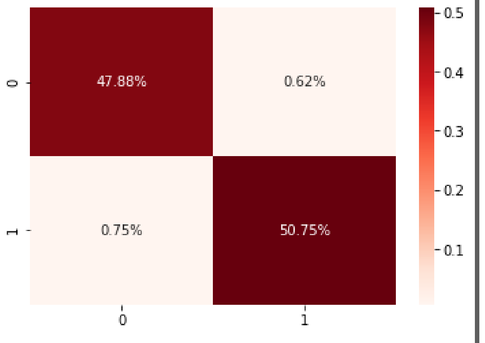
Predicting Fake News vs. Real News: *Continued on Page 4*

- Fitted an SVM and Logistic Regression Model based on TF-IDF word frequencies
- Achieved very high accuracy:
  - Logistic Regression: **99.12%**
  - SVM: **98.63%**
- However, may be due to quirks in the data
  - Sources for fake and real news are very uniform
- 1 for Fake, 0 For Real

Logistic Regression Confusion Matrix:



SVM Confusion Matrix:



Exploratory Data Analysis

INSTRUCTOR: ADAM ANDERSEN

DIGITAL HUMANITIES, SUMMER 2021

STUDENT: JASPER ZHOU

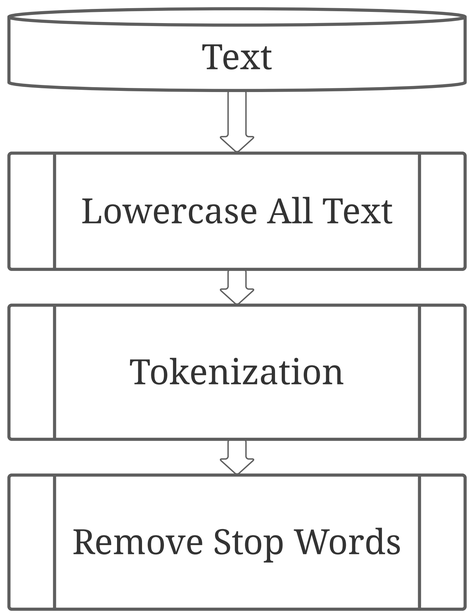
What is Exploratory Data Analysis?

- Exploratory Data Analysis aims to discover and summarize the dataset we are given
- Helps inform future analysis
- In some cases, EDA can help discover and change the questions we ask:
  - For example. the initial research questions for my project were:
    - Track how fake news has changed
    - Predict fake news and real news using NLP and machine learning Methods
  - However, this after EDA, I realized that dates for fake and real news were from January 1st 2015 to December 31st 2017, which does not give a large enough timeline to determine large trends in fake and real news.
  - Changed research questions to:
    - Can we use sentiment analysis tools to track differences between fake and real news?
    - Can we predict fake and real news?

What is Data Cleaning?

- Data cleaning transforms our data from its raw form to a usable form
- For example, if we were given the task of inputting receipt data from vendors across the world, we might encounter different ways money is denoted:
  - For example, in Spain, 1 Euro is expressed as 1,00 instead of 1.00 like we do in the United States
  - In addition, if we wanted to do analysis and compare sales, we may need to convert different currencies into a standard currency (e..g USD)
- Similarily, for text data, we need to clean our text to make it conducive to analysis
- For text analysis, the following steps are needed:
  - Lowecase all text
  - Tokenization
  - Removal of stop words

Data Cleaning Pipeline



Explanation of Pipeline

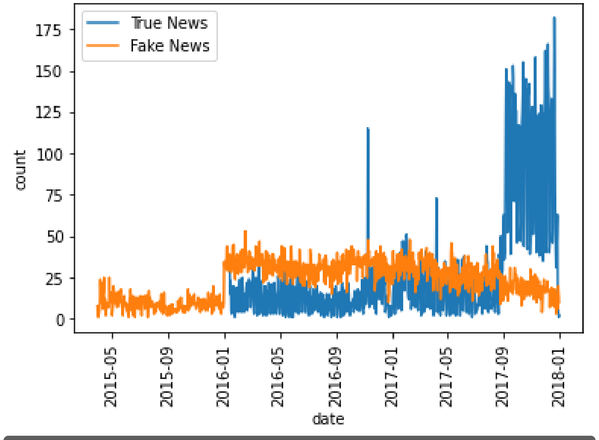
- **Lowercase all Text:** I converted all the text to lowercase in order to standardize the words and prevent mis-grouping of words in future analysis. For example, I do not wish the words "Important" and "important" to be classified as different words
- **Tokenization:** Tokenization splits a sentence into "tokens" or individual components. For example, the sentence "She sold seashells by the seashore" will be tokenized into a list containing ["she", "sold", "seashells", "by", "the", "seashore"]
- **Removal of Stop Words:** Stop words are words that are commonly used in sentences that don't offer "relevant" information and usually serve as connectives. For example, words such as "the" or "which". We wish to filter these out in so our EDA isn't cluttered by these phrases, which offer no insight into the content of an article

EDA Process:

- For my EDA, I did two things:
  1. Create a lineplot for the fake and real news: This will help examine what date ranges our data is from.

  2. WordCloud of fake and Real News: This will help examine the differences in topics between fake and real news

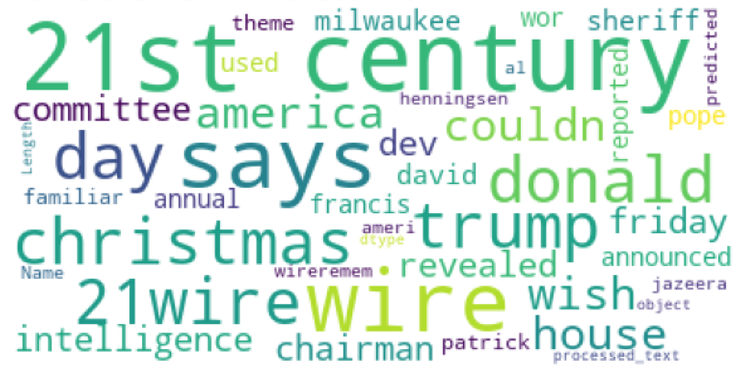
Fake and Real News Count



- We see that fake news spikes in 2016 and then slightly decreases until 2018
- We also see that true news significantly increases between September 2017 to January 2018
- I believe the spike in 2016 is attributed to Trump's presidential run
- True news may have increased due to Trump's first term in office. However, I have not confirmed this

Real News WordCloud

Fake News WordCloud



Conclusions from WordCloud

- We notice that the bulk of the real news is from Reuters and that the bulk of the fake news is from 21st Century Wire
  - Conclusion: The sources for this dataset are very homogenous
- Differences in topics discussed:
  - The different topics highlighted tracks with our intuition of conservative and liberal agendas
  - Because our fake news is from 21st Century Wire, a conservative blog, we see issues in the fake news that are commonly hot button issues for conservative beliefs. For example, "christmas", a commonly discussed topic of "the war on Christmas"
  - In addition, we can see the traditional conservative/liberal split on international affairs. For example, the fake news talks about "america", while the real news focuses on "nato" and "allies"
- High interest of real news in "conservative":
  - While real news is interested in conservative beliefs, it seems that fake news is less so
  - Perhaps real news is more reactionary than fake news

Sentiment Analysis

INSTRUCTOR: ADAM ANDERSEN

DIGITAL HUMANITIES, SUMMER 2021

STUDENT: JASPER ZHOU

What is Sentiment Analysis?

- Sentiment Analysis attempts to use computational methods to identify the tone of a text
- In our project, we are using the TextBlob package to determine the polarity and subjectivity of a sentence

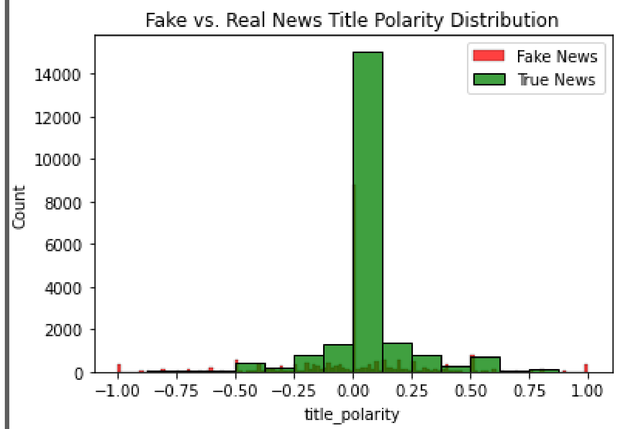
What is Polarity and Subjectivity?

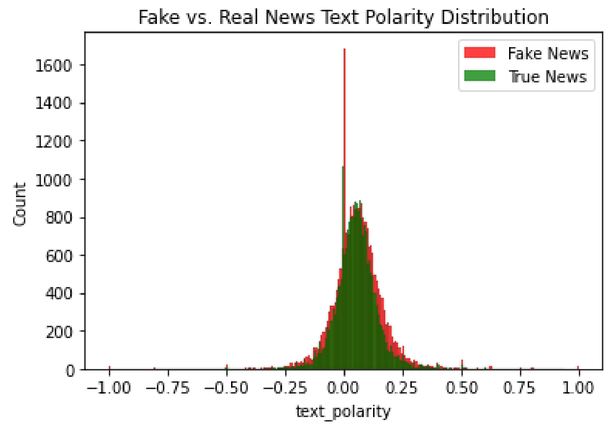
- **Polarity**: Polarity indicates how positive or negative a text is. In TextBlob, we are given a score from -1 to 1. A score of -1 indicates a sentence is very negative, while a score of +1 indicates a sentence is very positive
- **Subjectivity**: Subjectivity indicates how subjective a text is. Similarly, we are given a score from -1 to 1. A score of -1 indiciates a less subjective text, while a score of +1 indicates a very subjective text.
- TextBlob contains a dictionary of words and a score associated with each word. These words are matched to the sentence and then averaged to return a score for the entire sentence

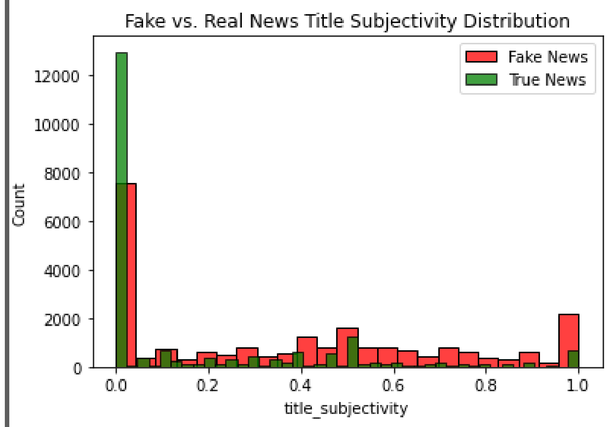
Sentiment Analysis Process:

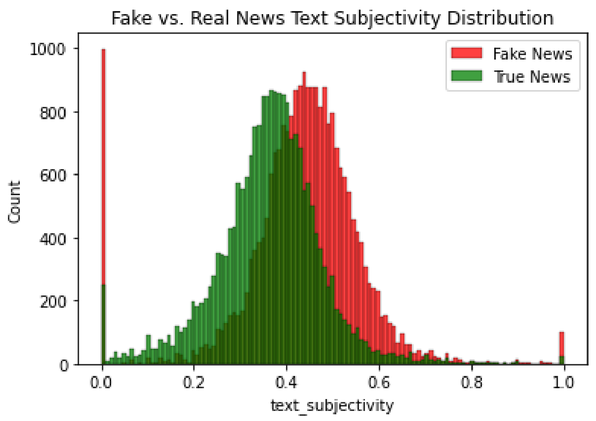
- I wanted to look at the distribution of polarity and subjectivity of the following:
  - Polarity of the title for fake and real news
  - Subjectivity of the title for fake and real news
  - Polarity of the text for fake and real news
  - Subjectivity of the text for fake and real news
- **Why does this matter?**
  - Doing so allows individuals to potentially use sentiment analysis as a way to augment our classification process
  - Reveals insights into the nature of fake and real news
  - Evaluate the suitability of TextBlob for sentiment analysis of fake and real news

Results of Sentiment Analysis









Conclusion of Sentiment Analysis:

- The distribution of the polarity and subjectivity of the titles are not significantly different
- This may be due to the shortness of the titles, which may not allow many matches between the words in the title and the words in the TextBlob dictionary
- Interestingly, the polarity of fake and real news are very similar.
- However, significant difference between the subjectivity of fake and real news
- On average, fake news is more subjective than real news.
- Standard Deviations of both are similar as well

Model Fitting

INSTRUCTOR: ADAM ANDERSEN

DIGITAL HUMANITIES, SUMMER 2021

STUDENT: JASPER ZHOU

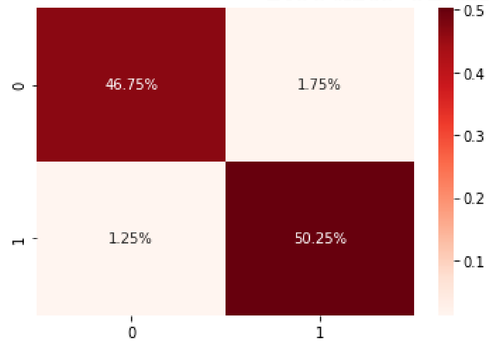
Model Fitting Process:

1. Vectorize our text into a TF-IDF matrix
2. Create a training, validation and test set of our vectorized data
3. Fit two models on training set: Linear SVM and Logistic Regression Model
4. Evaluate which performs the best on validation set
5. Take best model, so far, augment the TF-IDF matrix with subjectivity score
6. Fit using training set to create a new model
7. Evaluate all models on the test set

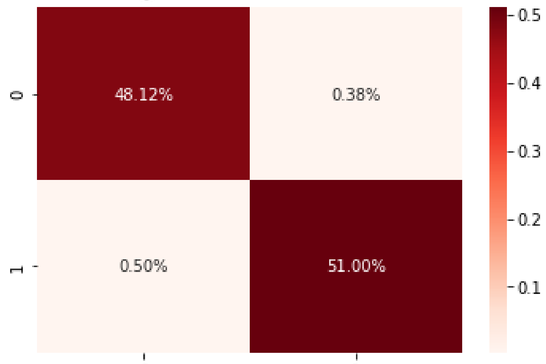
What is TF-IDF?

- TF-IDF stands for Term Frequency-Inverse Document Frequency
- Is a method to account for the frequency of words in a collection of documents
- Aims to weight words by importance
- TF-IDF consists of two parts: TF (Term Frequency and IDF (Inverse Document Frequency)
- TF-IDF = TF x IDF, where:
  - TF stands for the term frequency, which could be a simple count of each word in each document
  - IDF stands for Inverse Document Frequency, which is the inverse of the fraction of number of documents where term t appears among the total documents
  - Intuitively, a high TF-IDF score

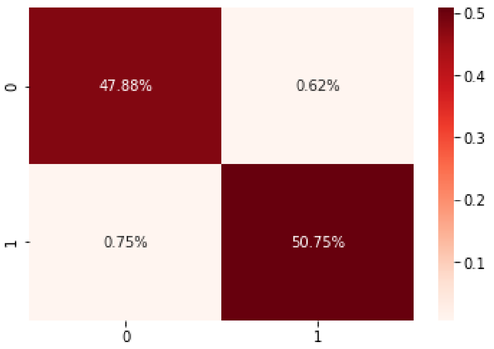
Results of the Model Fitting Process:



Confusion Matrix for Augmented SVM



Confusion Matrix for Logistic Regression



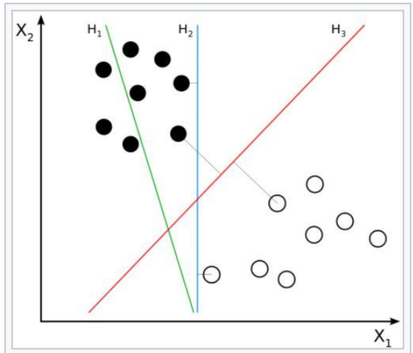
Confusion Matrix for SVM

Conclusions from the Model Fitting Process:

- Accuracy of the TF-IDF only Logistic Regression Model: 99.12%
- Accuracy of the SVM of the TF-IDF only SVM Model: 98.63%
- Accuracy of the Augmented SVM: 97.00%
- Interestingly, augmenting the TF-IDF matrix with subjectivity scores lowered the accuracy
- Unusually high accuracy of the TF-IDF only Logistic Regression
  - I believe this is due to the fact the sources were very homogenous
  - As a result, a very high weight was placed on the specific source
- For future work, two approaches are possible
  - Keep the current data, find all the sources, then remove all the text which mention the source in the article
  - Conduct own web scraping and follow the approach above. With own web scraping, can know the exact source and remove mentions of the source in the text

What is SVM?

- Support Vector Machines
- Classifies data into binary categories (0 or 1)
- Aims to find a hyperplane (in 2D this would be aline) where the average distance between the hyperplane and each point is the largest



H_1 does not separate the classes. H_2 does, but only with a small margin. H_3 separates them with the maximal margin.

Source: Wikipedia

What is Logistic Regression?

- Classifies data into binary categories
- Aims to find a linear model such that the log odds of an event category occuring (1) is maximized
- Where odds are defined as p/(1-p) where p represents the probability an event occurs
- Then, a threshold is defined (default = 0.5)
- If p>= 0.5, then classify as 1
- if p < 0.5, then classify as 0
- Source: Wikipedia