CS88 Fall 2022
Logisitic Regression for Fake News Detection

Brooke Coneeny and Sydney Levy

December 7th 2022

1 Introduction

In recent times news sources are vastly available, either through television, the internet, or social media. Because such a large audience is reading these stories, it is critical that their integrity is being inspected. The term ‘Fake News’ is used to refer to news which is false or presents misleading information. We will refer to news that is not considered to be fake news as ‘True News.’ Fake news articles contain no verifiable facts, sources, or quotes. The vast way in which these stories are being published allows fake news stories to circulate at a rapid pace, which is why it is important that they are detected both quickly and accurately. Many people believe that any source they find online is credible, meaning that it is especially important to be able to classify fake news sources to ensure that people are not misled.

The prevalence of fake news online goes against a couple of the security goals which we have covered in this course. Many fake news sources are widely available and spread quickly throughout social media. Thus, while it is an important security goal to promote availability, we want to ensure that available news sources also have integrity and accountability. One of the main security goals of any system is integrity, the prevention of unauthorized changes and the trustworthiness of data/resources. These fake news reports can often include quotes which were changed without permission and / or references to untrustworthy data sources. Similarly, these reports often do not uphold accountability, the ability to provide evidence that a specific action occurred. Since many fake news stories can be posted anonymously or under false names, there is a lack of accountability for who wrote the story. Additionally, because the sources used in these articles cannot be trusted, there is no reliable evidence that any specific action referenced in the report actually occurred.

2 Methods

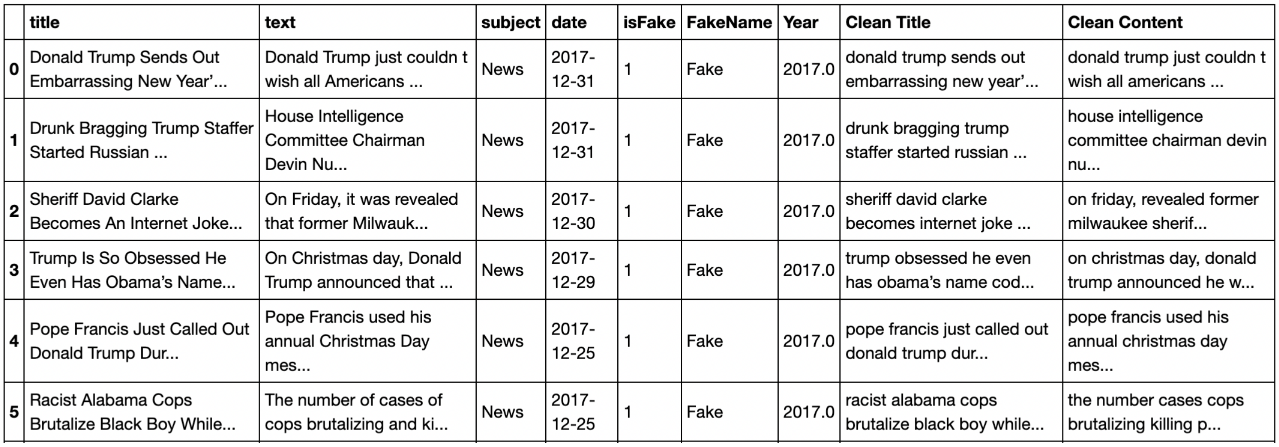
2.1 About the Data

The dataset we used for this project can be found at: `https://www.kaggle.com/code/maxcohen31/nlp-fake-news-detection-for-beginners/data`. It consists of 23,481 Fake News articles (labeled as 1) and 21,417 True News articles (labeled as 0). For each article, the dataset contains information on its title, subject, content, and date published.

2.2 Cleaning the Data

Before beginning our exploratory data analysis, we cleaned the dataset. The first thing we did was create an additional column called 'Year' which contains only the year the article was written and not the entire date. We then went on to remove popular stop words from both the titles and text. The stop words we removed include: stop, the, to, an, and, so, a, in, it, is, I, that, had, on, for, were, was, at, of, in, are, with, by, from, this, as, not, be, and will. Specifically, we removed the first 10 stop words from each title and the first 50 stop words from the article content. These cleaned article components can be found in the 'Clean Title' and 'Clean Content' columns of our dataset. We chose to remove stopwords so that our analysis does not focus on low-level features of the text, but instead relies on the information provided in the title or content of the articles. A sample of this cleaned data table can be found in Figure 1.

Figure 1: Sample of Data

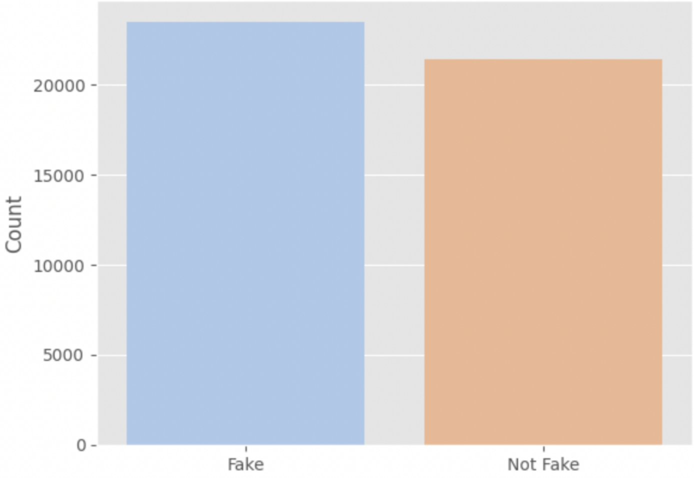


	title	text	subject	date	isFake	FakeName	Year	Clean Title	Clean Content
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	2017-12-31	1	Fake	2017.0	donald trump sends out embarrassing new year'...	donald trump just couldn t wish all americans ...
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	2017-12-31	1	Fake	2017.0	drunk bragging trump staffer started russian ...	house intelligence committee chairman devin nu...
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	2017-12-30	1	Fake	2017.0	sheriff david clarke becomes internet joke ...	on friday, revealed former milwaukee sherif...
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	2017-12-29	1	Fake	2017.0	trump obsessed he even has obama's name cod...	on christmas day, donald trump announced he w...
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	2017-12-25	1	Fake	2017.0	pope francis just called out donald trump dur...	pope francis used his annual christmas day mes...
5	Racist Alabama Cops Brutalize Black Boy While...	The number of cases of cops brutalizing and ki...	News	2017-12-25	1	Fake	2017.0	racist alabama cops brutalize black boy while...	the number cases cops brutalizing killing p...

2.3 Exploratory Data Analysis

The first element of our dataset we wanted to examine was the proportion of Fake News to True News available. In order for our model to best learn the data, there needed to be and almost equivalent amount of each type of article. As we can see in Figure 2, there did seem to be a nearly equivalent proportion of Fake to True news articles (with greater than 20,000 articles of each type).

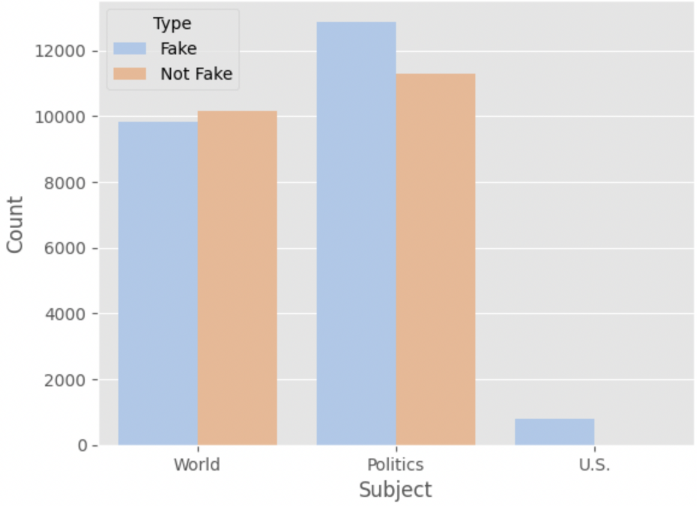
Figure 2: Percent of Fake News in Dataset



Given that both article types were categorized by subject, we wanted to analyze which subjects are most prevalent in Fake News articles. Before beginning this exploration, we needed to further clean our dataset. Originally, there were different subject names for the same topic depending on whether it came from the Fake dataset or the True Dataset. For example, one such redundancy was subjects being labeled as ’politicsNews’ and ’Politics’. Condensing the subjects led to the following categories: Politics, World, and U.S.

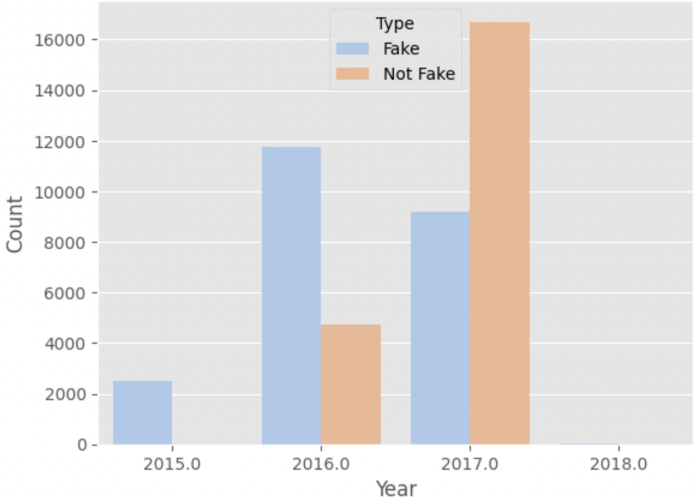
Figure 3 summarizes how often each of these subjects appear in Fake or True News. The blue bars represent Fake News articles, and the orange bars represent True News articles. We can see below that more political news stories were fake, although it was not a vast difference between the two. For world news, more stories were True, but once again there was not a large difference in the proportion. Lastly, it can be noticed that all U.S. news articles in our dataset were categorized as Fake, which will not be ideal when it is time for our model to learn the data.

Figure 3: News by Subject



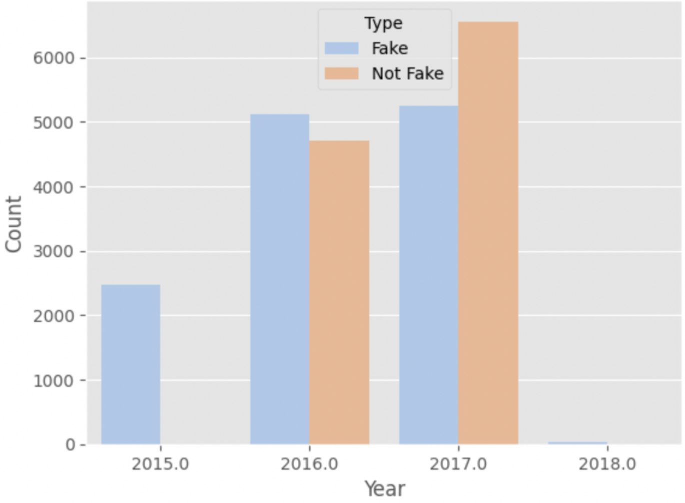
This dataset includes articles from an election year (2016) and we therefore wanted to see the differences in proportion between Fake and Real News over the time frame. The relationship is visualized in Figure 4 below. As we can see, a large majority of articles published in 2016 were Fake News. This is due to competing political figures attempting to boost their campaigns and sink others. After the election year it can be seen that the proportion of Fake and Real News swapped from 2016.

Figure 4: News by Year



Given there is an election year included in our dataset, we wanted to investigate whether there was a difference in the proportion of political news that was fake in election years vs. non-election years. As can be seen in Figure 5, during the election year there was much more Fake news than real news. Further, in 2017 right after the election, there was much more True news than Fake news. Thus, it seems especially important to classify and take note of Fake news sources during election years.

Figure 5: Political News by Year in Dataset



3 Results

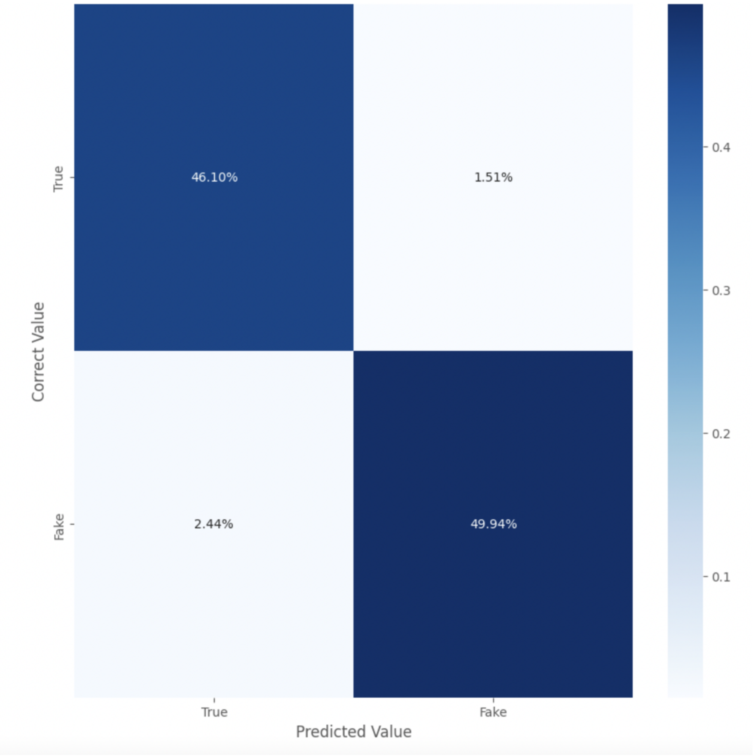
3.1 Logistic Model

We decided to use a Logistic Regression to predict whether an article is Fake or True News. A Logistic model is a classification model, meaning it is used when the dependent variable is categorical (Fake or True News), making it a good fit for our research.

3.1.1 Model 1: News Titles

The first Logistic Model we created predicts Fake News based on the title of the article (with stopwords removed). Before training our model, we broke our data into training (80%) and testing (20%) sets. After the training was complete, our model performed with 96% accuracy on the testing set. Figure 6 demonstrates which types of errors the model makes. Although this error is small, the majority of it comes from Fake News articles which the model predicted to be Real News (2.44% of the time). The model makes errors predicting that True news is Fake 1.51% of the time.

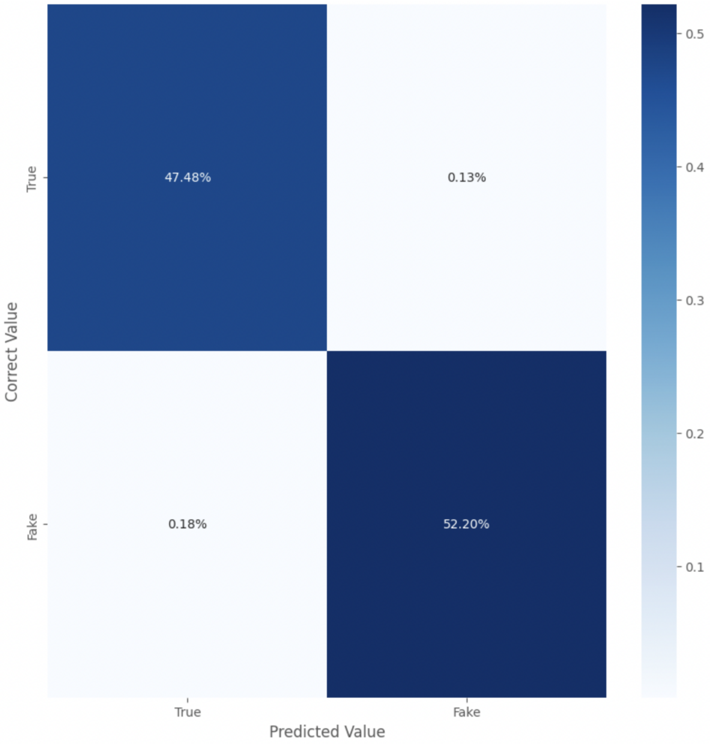
Figure 6: News Titles: Logistic Model Confusion Matrix



3.1.2 Model 2: News Content

The next Logistic Model we created predicts Fake News based on the content of the article (with stopwords removed). Once again, we broke our data into training (80%) and testing (20%) sets. After the training was complete, the model performed with 99% accuracy on the testing set, therefore improving from Model 1. Below Figure 7 demonstrates which types of errors the model makes. This suggests that the content of the article is more useful than just the title for determining whether the article is Fake or not. The model predicts True articles to be fake 0.13% of the time and predicts Fake articles to be True 0.18% of the time.

Figure 7: News Content: Logistic Model Confusion Matrix



3.2 Word Frequency Analysis

Given the title of news articles is what circulate most quickly online, we decided to analyze if there were specific words that were more prevalent in the titles of Fake news stories. We created bar charts to visualize the 15 most prevalent words in the titles of all the Fake and True news stories (after stopwords were excluded). As we can see below in Figure 8 and Figure 9, ”Trump” was the most frequently used word in both types of articles, which could cause confusion for our models. It can also be seen in Figure 8 that articles which specify they include a video are very prominent in Fake News but not even in the top 15 most frequent words for Real News. This will make it much easier for our model to predict the type of article correctly based on whether it includes a video. Lastly, it can be seen that True articles contains more political words such as ”u.s.”, ”house”, ”senate”, ”Russia”, ”Korea”, ”court”, and ”republican”. This is likely because True News articles focus on providing evidence and context about political matters, whether this be information on countries such as Russia / North Korea or about rulings in the House, Senate or Supreme Court.

Figure 8: Top 15 Words in Fake News Title

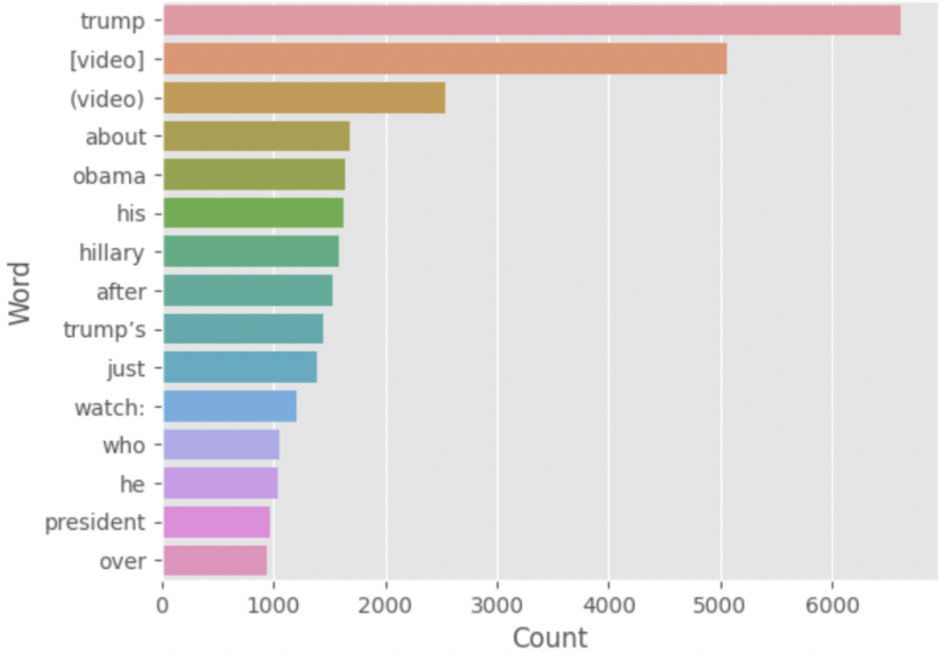
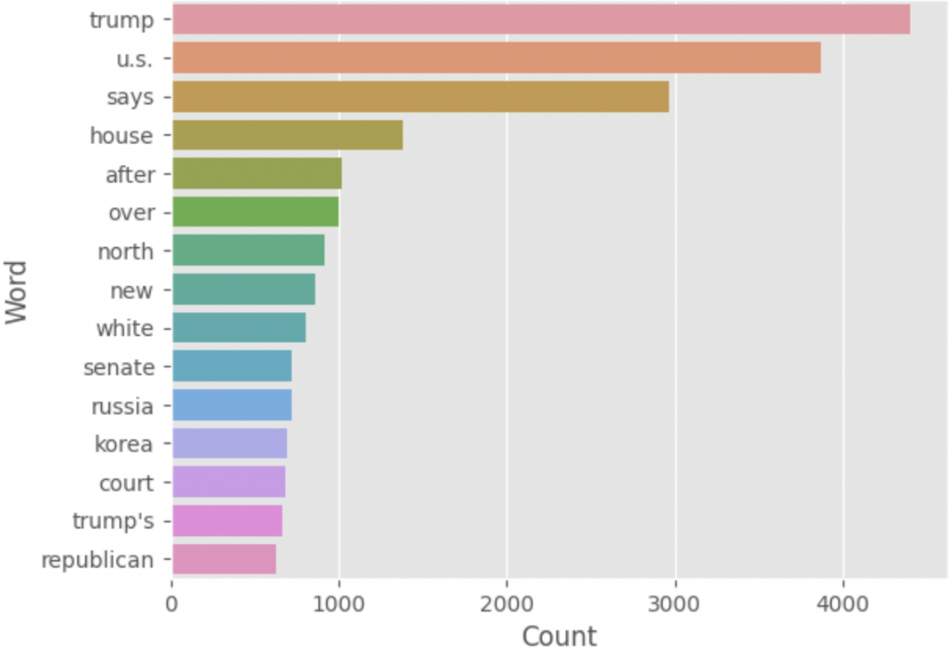


Figure 9: Top 15 Words in True News Title



4 Conclusions

Our project accurately trains Logistic Regression models to identify Fake News articles. The first model predicts the type of news based on the article's title with 96% accuracy. The second model predicts the type of news based on the article's content with 99% accuracy.

As for the articles which were misclassified by the models, the majority represented Type II errors (False Negatives). One possible explanation is that the model was being confused by the prevalence of the word "Trump" in both types of articles. Going forward, we would like to create a model which predicts the type of article based on both the title and the content.

5 References

Link to Dataset: https://www.kaggle.com/code/maxcohen31/nlp-fake-news-detection-for-beginners/data

Link to Code Referenced for Logistic Model: https://www.kaggle.com/code/paramarthasengupta/fake-news-detector-eda-prediction-99

Link to Code Referenced for Analysis of Word Frequencies: https://www.kaggle.com/code/sreshta140/is-it-authentic-or-not