Introduction and Problem Statement

Do you trust all the news you hear from social media?

All news are not real, right?

How will you detect fake news?

The rapid spread of fake news has become a major issue worldwide. The spread of false and misleading news has led to significant social and economic consequences, impacting industries from finance to healthcare. In this context, this project will explore opportunities that exist the identify fake news as soon as it is available on social media to limit its influence on people, communication, and to prevent confusion. The criteria for success would be a model that accurately identifies patterns in news that a considered to be fake. The scope of the solution space is using sample data for detecting fake articles; however, the process can be used to detect fake tweets or anything similar. A major constraint would be the amount of data that is available on the internet. Privacy concerns would be another constraint as it may limit access to data. Stakeholders include any user of the internet.

What is Fake News?

A type of yellow journalism, fake news encapsulates pieces of news that may be hoaxes and is generally spread through social media and other online media. This is often done to further or impose certain ideas and is often achieved with political agendas. Such news items may contain false and/or exaggerated claims, and may end up being viralized by algorithms, and users may end up in a filter bubble.

Data

The data was obtained from Kaggle Data of Fake and True news dataset. The dataset consists of two separate csv files from various news sources and are already labelled as fake and true. The dataset consists of 4 columns. The news is categorized into subjects and is ordered by date.

A label column was added to identify the dataset and then was merged as one pandas file as shown below:

	title	text	subject	date	label
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	TRUE
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	TRUE
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	TRUE
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	TRUE
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	TRUE

Tools Used

Pandas: Data loading, manipulation, wrangling and

Scikit learn: Libraries for text feature extraction, vectorizer, metrics, classifier models and cross validation

NLP: stop words removal, text processing, n-gram analysis

Matplotlib and Seaborn: Data visualization

Data Wrangling

The data was read using pandas. The dataset consists of 2 csv files each containing news that are already labelled as true and fake. Each file consists of 4 columns. The news is categorized into subjects and is ordered by date. After reading the data, we created a column named label with two labels TRUE and FAKE. This will be our prediction variable later on.

The describe method was used to get descriptive statistics about the dataset. From the table below we can see that there are approximately 45000 news with 39000 being unique. There are 14 unique titles and 11272 subjects.

	title	text	subject	date	label
count	44898	44898	44898	44898	44898
unique	38729	38646	8	2397	2
top	Factbox: Trump fills top jobs for his administ...		politicsNews	December 20, 2017	FAKE
freq	14	627	11272	182	23481

Missing values were checked and there were none. In order for the natural language processing (NLP) model to analyze text data, the data had to be cleaned and prepared the into a machine-readable format. The first step in this data cleaning process is to remove tags and numbers as they do not convey any inherent information for the prediction. Also, contractions like ain't are difficult for models to read and so a function was written to convert any contractions into its expanded form. Then, any additional white spaces were scraped and converted the text into lowercase.

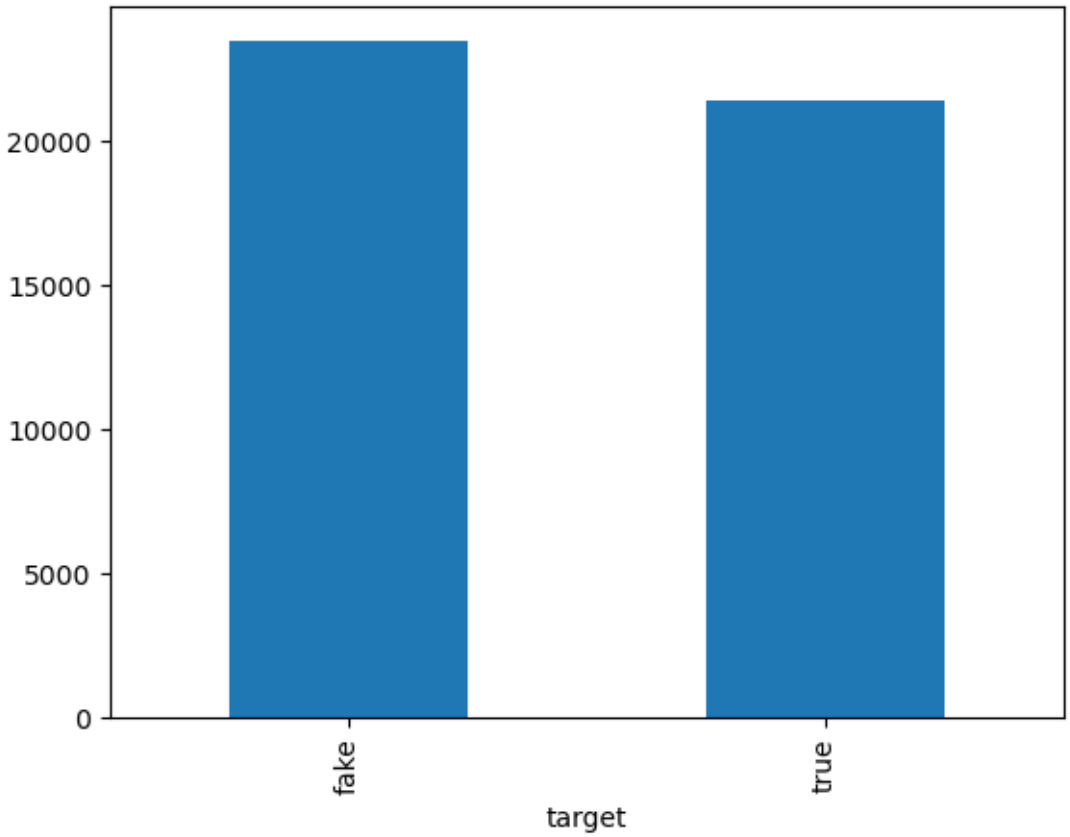
The next step was to apply lemmatization with stopwords removal. Lemmatization is the process of converting a word into its root form. For example, having and have are essentially the same word except for the tense. With lemmatization, having is converted to have. We also remove stopwords such as 'and' and 'the'. Stopwords removal depends on the application of NLP. In this case, large articles were being analyzed which contained a large number of stopwords. If these were not removed, they would bias the model and make predictions useless. After that, any remaining punctuation was removed. Finally, we have clean data and are ready for analysis.

	title	text	subject	date	label
0	as us budget fight loom republicans flip fisc...	washington reuters the head conservative re...	politicsnews	December 31, 2017	TRUE
1	us military accept transgender recruit monday ...	washington reuters transgender people allow...	politicsnews	December 29, 2017	TRUE
2	senior us republican senator let mr mueller job	washington reuters the special counsel inve...	politicsnews	December 31, 2017	TRUE
3	fbi russia probe helped australian diplomat ti...	washington reuters trump campaign adviser g...	politicsnews	December 30, 2017	TRUE
4	trump want postal service charge much amazon ...	seattlewashington reuters president donald ...	politicsnews	December 29, 2017	TRUE

Exploratory Data Visualization

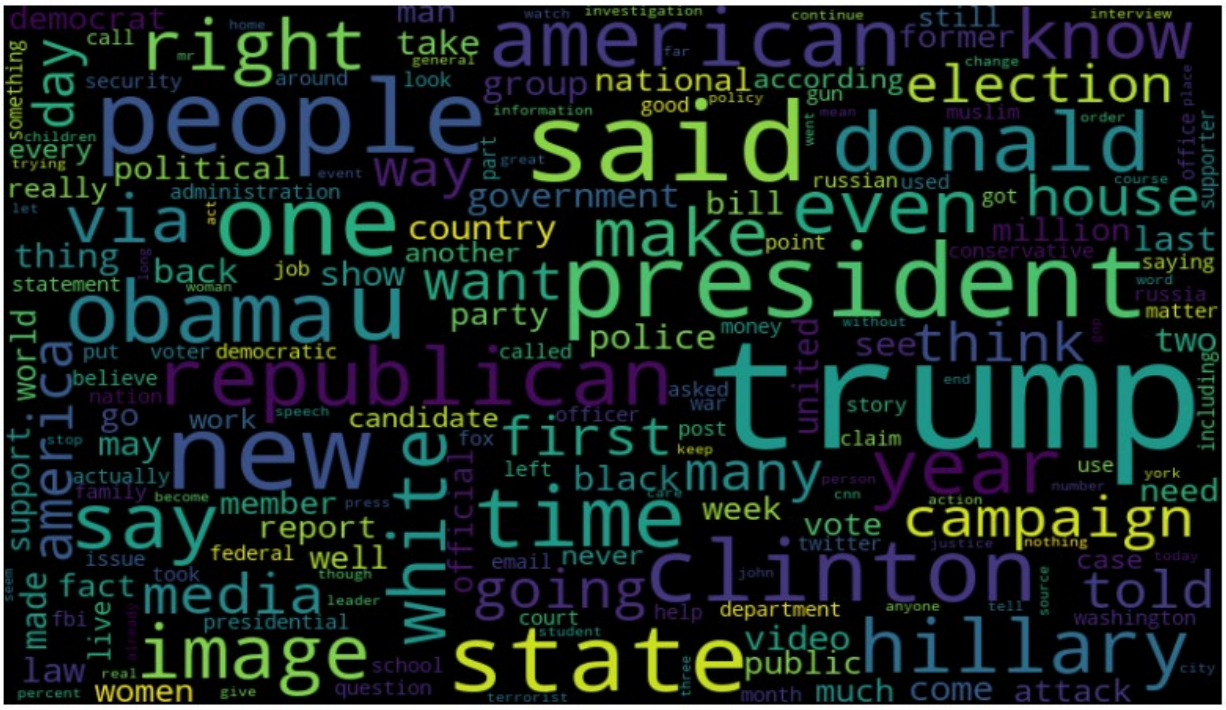
The cleaned data was read using pandas. The label was mapped as TRUE into 1 and FAKE into 0. The distribution of TRUE and FAKE labels in the dataset was used to verify if the dataset is balanced.

As noted in graph below, there was a roughly even balance between the two classes. This means that there was not a need to run any undersampling or oversampling.

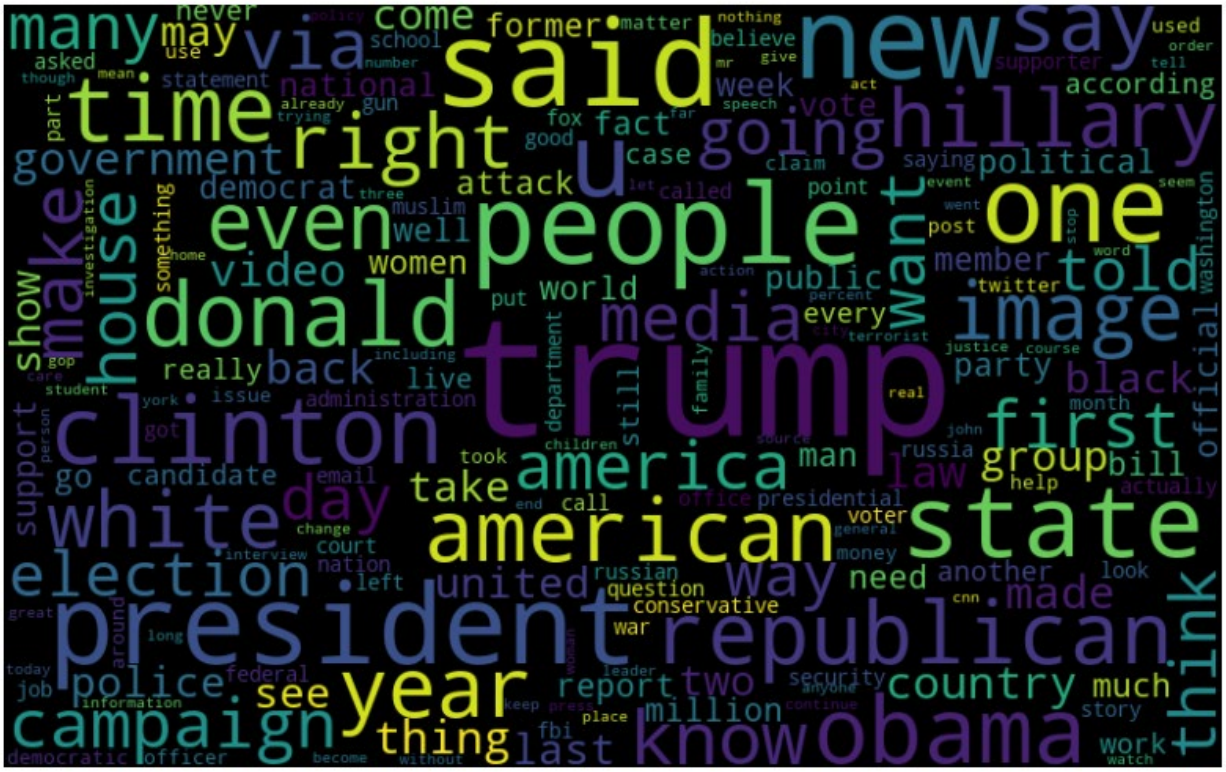


A bar chart was plotted of the number of articles per subject. There were three (3) predominant articles per subject and the remaining subjects were from five (5) other articles.

World clouds were created to see the words that were most common in each of the categories. Trump, said, people, president, and would seems to be the most prominent words in Fake News.



Looking at the wordcloud of true news, the most prominent words were said, Trump, us, would, and Reuters.



Feature Engineering

In order to perform feature extraction, the scikit learn package TF-IDF transformer was used. **TF-IDF** stands for Term Frequency — Inverse Document Frequency and is a statistic that aims to better define how important a word is for a document, while also considering the relation to other documents from the same corpus.

This is performed by looking at how many times a word appears into a document while also paying attention to how many times the same word appears in other documents in the corpus. So, then **TF-IDF is a score** which is applied to every word in every document in the dataset. And for every word, the TF-IDF value increases with every appearance of the word in a document but is gradually decreased with every appearance in other documents. This score is then fed into the algorithm.

Algorithms and ML Model

Again, scikit learn library was imported to use a bunch of different algorithms to compare the accuracy and precision of the predictions. The models used in this learning were the Passive Aggressive Classifier, Decision Tree Classifier, and Random Forest Classifier. From the three (3) models, it is observed that all of them have very few misclassifications. The accuracy scores were reported as follows: Passive Aggressive Classifier (95.55%), Decision Tree Classifier (88.15%), and Random Forest Classifier (90.73%).

The top model was the Passive Aggressive Classifier based on the data and the confusion matrix.

Conclusion

In this project, the objective was to predict fake and true news using NLP and machine learning methods. There was success in developing a Passive Aggressive Classifier model that is capable of predicting TRUE and FAKE news with and accuracy score of 95.55%.