FAKE NEWS Detection

Nusrat Islam

Problem Statement

- In this era of social media, it has become a huge difficulty to identify fake news from true news.
- From democracy to the global economy everything is affected largely by false information. In that context, it is really important to identify fake news as soon as it is available on social media to limit its influence on people.

Data

- The data was collected from Kaggle
- The dataset consists of 2 csv files each containing news that are already labelled as true and fake

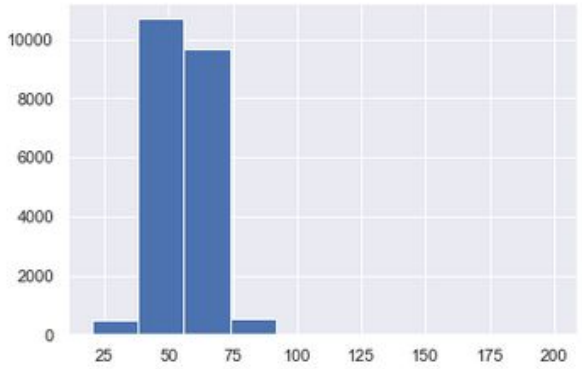
	title	text	subject	date	label
count	44898	44898	44898	44898	44898
unique	38729	38646	8	2397	2
top	Factbox: Trump fills top jobs for his administ...		politicsNews	December 20, 2017	FAKE
freq	14	627	11272	182	23481

Data Cleaning

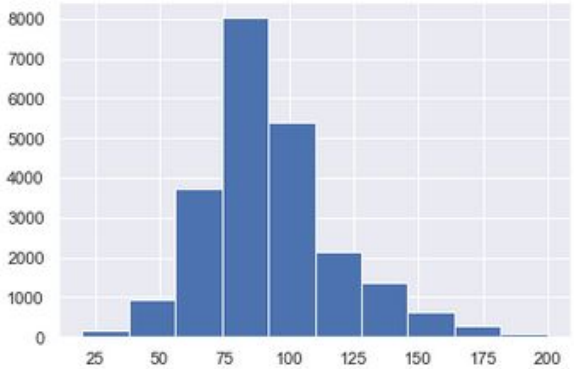
- Descriptive statistics was used to check for missing values
- Data cleaning was performed to remove tags, white spaces, numbers and contractions
- Using NLTK library lemmatization was used to convert all words to its root. This helps to identify similar words
- Stopwords e.g, the, a, of, etc. were removed as well

Exploratory data Analysis

- In this section we compared the length of text in the TRUE and FAKE news labels.
- We also looked at top 25 n-grams and wordcloud to identify common theme in each category and check for any underlying patterns.



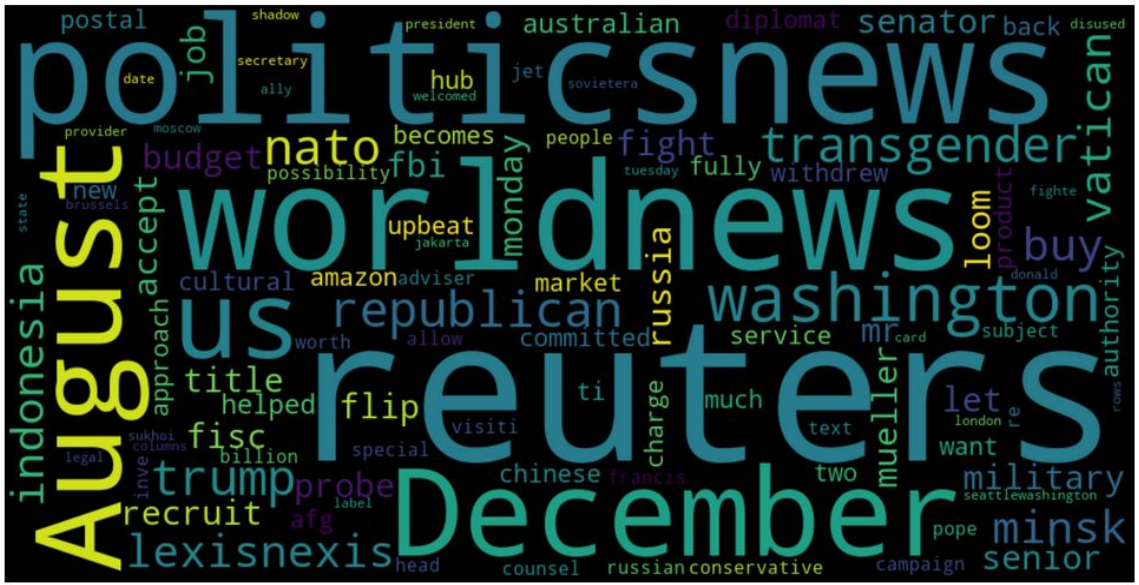
Number of characters in TRUE title



Number of characters in FAKE title

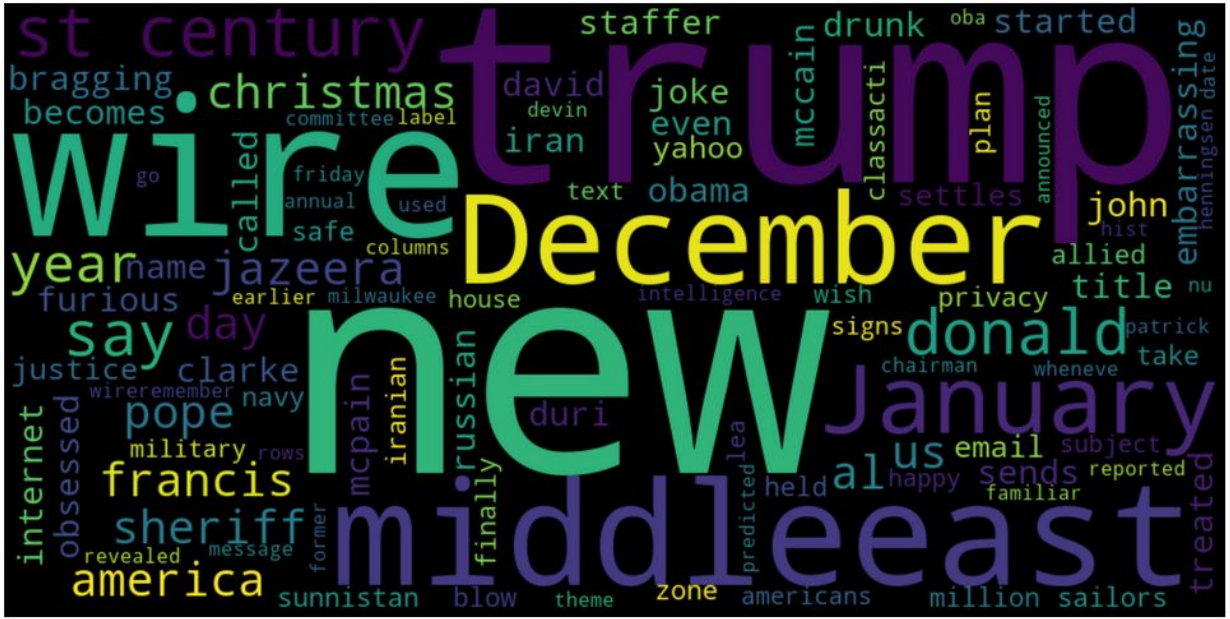
Wordcloud TRUE News

Reuters, worldnews, politics news,August, December, us and republican seemed to be the most prominent words in TRUE news.



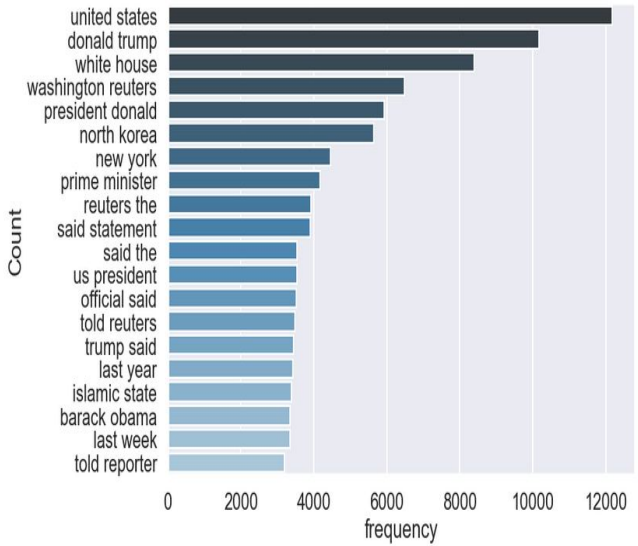
Wordcloud FAKE News

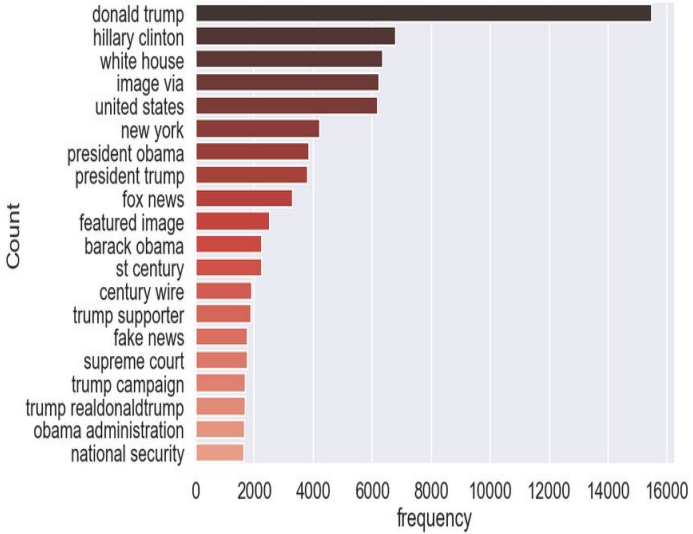
Looking at the wordcloud of fake news we find President Trump, wire, December, January, middle east being more prominent.



N-gram analysis

- High dimensional data is visualized by converting them into vectors
- The news mostly contains political content and president Trump is most common across both labels.

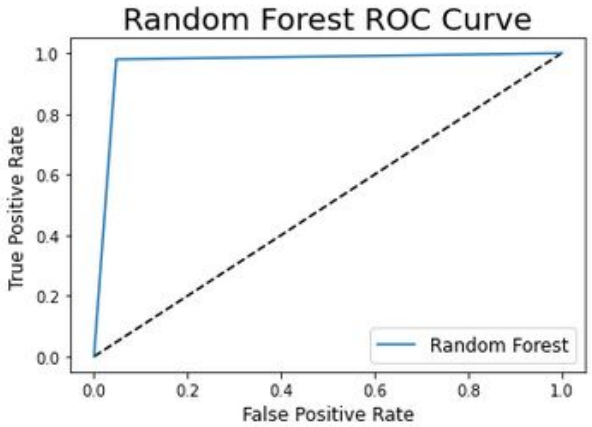


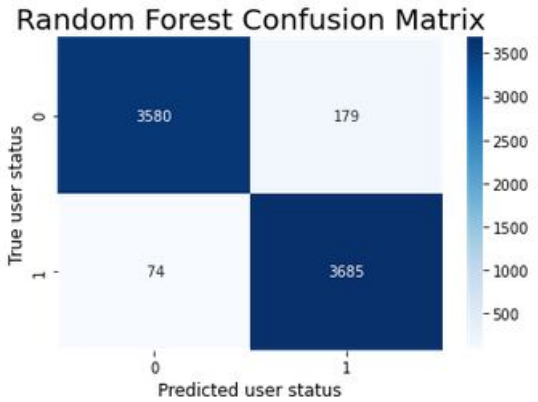


Modeling and Results

The models used in this learning are Multinomial Naive Bayes, Random Forest, Logistic Regression, Stochastic Gradient Descent.

From the 4 models it is observed that the exception of Naive Bayes, all of them have very few mis-classifications. The ROC_AUC score is 0.964 which is exceptionally well for text classification.





Conclusion

- In this project we aimed to predict fake and true news using NLP and machine learning methods.
- We were successful in developing a Random Forest model that is capable of predicting TRUE and FAKE news with and ROC-AUC score of 0.964.
- Although the accuracy is high, the data is heavily biased towards the US 2016 Presidential Election and so will not be able to generalize well.