

FAKE NEWS DETECTION PROJECT

Submitted by:

DEBANTI ROY

**ACKNOWLEDGMENT**

I, Debanti Roy, would like to convey my sincere gratitude to DataTrained Academy and Flip Robo Technologies for giving me this opportunity to do this project. I would like to thank all mentors and SME’s for extending their support all through the process which helped me complete this project.

**E-source:**

<https://en.wikipedia.org/wiki/Fake_news>

<https://library.gannon.edu/c.php?g=728528&p=5206339>

**INTRODUCTION**

* Business Problem Framing

Fake news has become one of the biggest problems of our age. It has serious impact on our online as well as offline discourse. One can even go as far as saying that, to date, fake news poses a clear and present danger to western democracy and stability of the society.

* Conceptual Background of the Domain Problem

Fake news's simple meaning is to incorporate information that leads people to the wrong path. Nowadays fake news spreading like water and people share this information without verifying it. This is often done to further or impose certain ideas and is often achieved with political agendas.

For media outlets, the ability to attract viewers to their websites is necessary to generate online advertising revenue. So it is necessary to detect fake news.

* Review of Literature

Fake news is false or misleading information presented as [news](https://en.wikipedia.org/wiki/News). Fake news often has the aim of damaging the reputation of a person or entity, or making money through [advertising](https://en.wikipedia.org/wiki/Advertising) revenue.

* Motivation for the Problem Undertaken

Fake news is a term that refers to news outlets that pretend to be real but are actually fake. Fake news includes false stories about all sorts of topics that are crafted in order to appear real and get traffic and social shares on Facebook and Twitter.

Our motive is to detect them and make users aware that they are coming across a fake news.

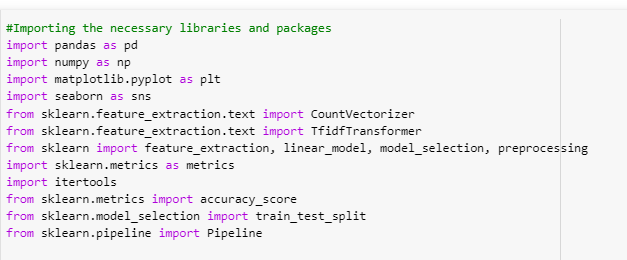
**Analytical Problem Framing**

* Data Sources and their formats

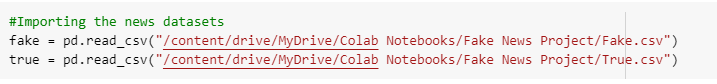
The datasets are provided by Fliprobo. There are two datasets one for fake news and one for true news. Both the files are in CSV format. In true news, there is 21417 news, and in fake news, there is 23481 news.

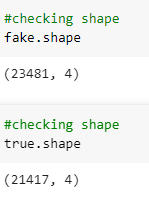
* Data Preprocessing Done
* **Importing the necessary libraries and packages**

First we have imported the necessary libraries.



Then we have imported our dataset which was in CSV format and printed the shape of the dataset, i.e., the total rows and columns.

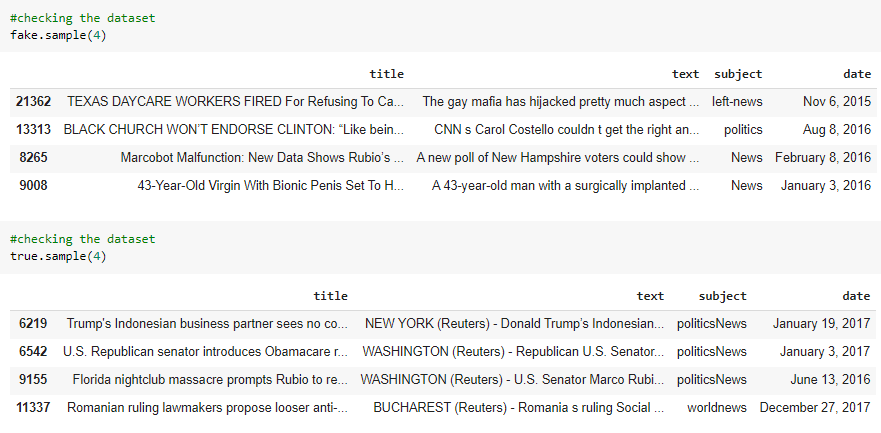




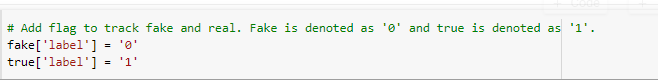
We can see the fake dataset has 23481 rows and 4 columns and true dataset has 21417 rows and 4 columns.

* **EDA**

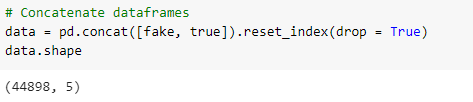
As a part of the Exploratory Data Analysis or EDA we have printed 5 rows as sample to get a first view of our dataset.

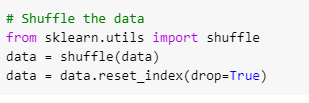


We added a separate column ‘label’ as flag to identify true and fake news.

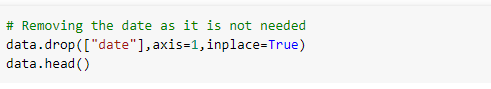


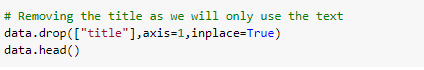
We have then concatenated both the fake and true news dataframes into one and then shuffled the data.



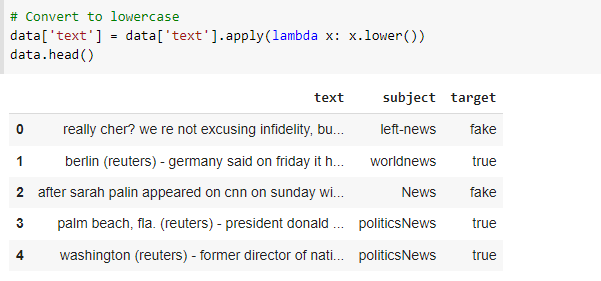


We have dropped that date and title columns as they were not necessary for out detection.

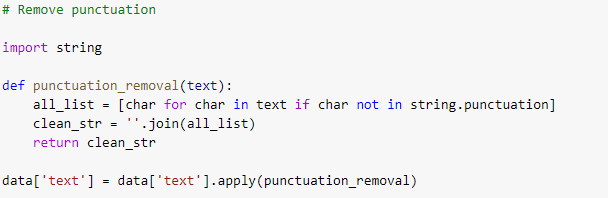




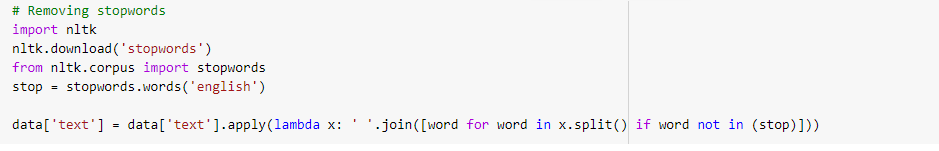
Next, we have converted the data into lower case.



We have also removed any punctuation marks from the dataset.



Next, we proceed to remove the stopwords from our dataset.



* Hardware and Software Requirements and Tools Used

Hardware required:

* 1. Processor: core i5 or above
  2. RAM: 8 GB or above
  3. ROM/SSD: 250 GB or above

Software required:

* 1. Anaconda 3- language used Python 3
  2. Microsoft Excel Libraries: The important libraries that I have used for this project are below:

*import numpy as np*

It is defined as a Python package used for performing various numerical computations and processing of the multidimensional and single dimensional array elements. The calculations using Numpy arrays are faster than the normal Python array.

*import pandas as pd*

Pandas is a Python library that is used for faster data analysis, data cleaning and data pre-processing. The data-frame term is coming from Pandas only.

*import matplotlib.pyplot as plt and import seaborn as sns*

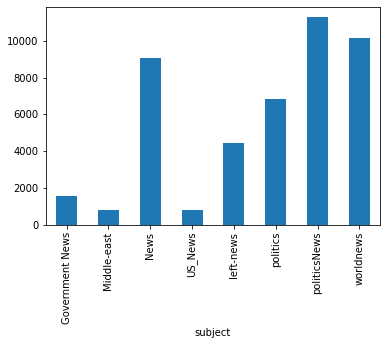
Matplotlib and Seaborn acts as the backbone of data visualization through Python.

**Matplotlib**: It is a Python library used for plotting graphs with the help of other libraries like Numpy and Pandas. It is a powerful tool for visualizing data in Python. It is used for creating statical interferences and plotting 2D graphs of arrays.

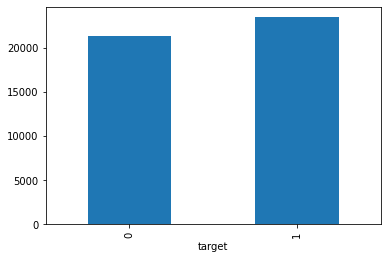
**Seaborn**: It is also a Python library used for plotting graphs with the help of Matplotlib, Pandas, and Numpy. It is built on the roof of Matplotlib and is considered as a superset of the Matplotlib library. It helps in visualizing univariate and bivariate data.

**Model/s Development and Evaluation**

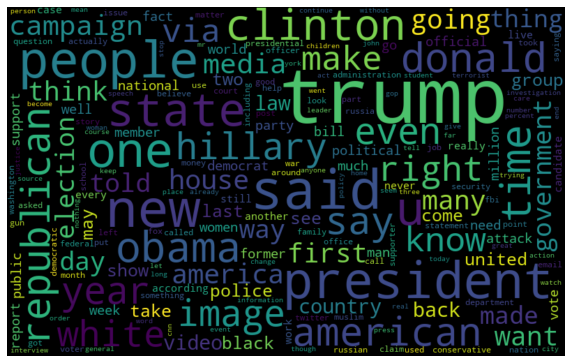
* Identification of possible problem-solving approaches (methods)
* I have used “.drop()” function to drop unwanted entries in the columns.
* Used concat and shuffle functions to concatenate fake and true dataframes into one and then shuffle them.
* Performed Lowercase conversion, stopwords removal, punctuation removal, tokenization, etc.
* Performed univariant and bivariate analysis using seaborn and matplotlib.
* Plotted wordcloud for both fake and true news.
* Visualizations

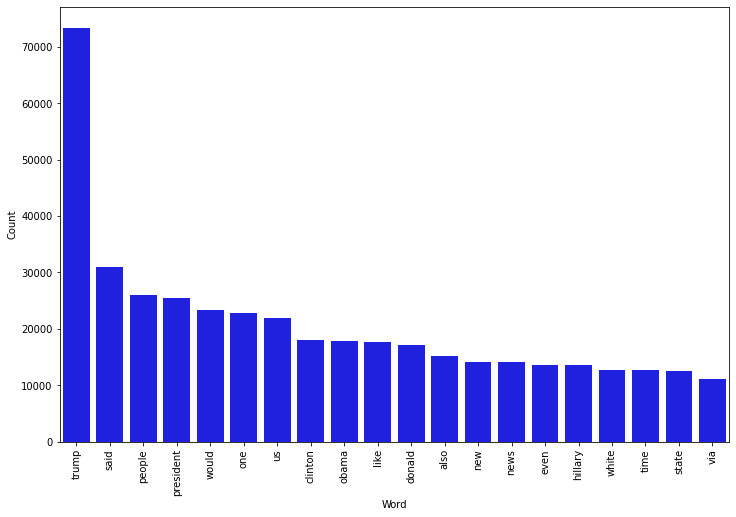


News count on the basis of subject. There are more political news than any other news.



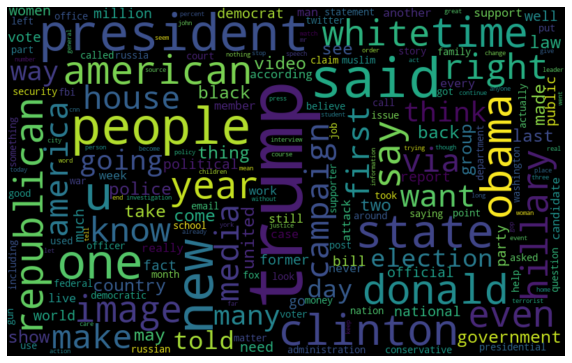
“1” denotes fake and “0” denotes true. We can see there are more fake news.

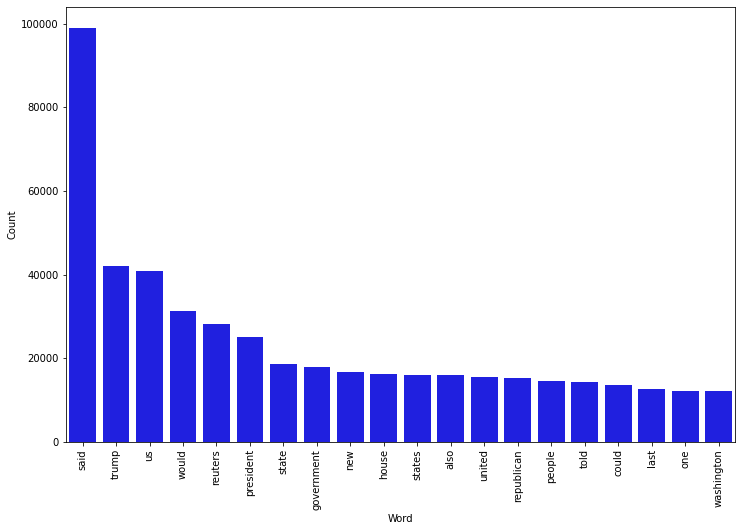




Observation:

The fake news word cloud shows that words like trump, said, people, president, would, etc. are the most used.





Observation:

The true news word cloud shows that words like said, trump, us, would, reuters, president etc. are the most used.

* **Interpretation of the Results**

|  |  |
| --- | --- |
| **Model Name** | **Model Accuracy** |
| **Naïve Bayes** | **95.2%** |
| **Logistic Regression** | **98.99%** |
| **Decision Tree** | **99.57%** |
| **Random Forest** | **99.37%** |
| **SVM Classifier** | **99.6%** |

* **Conclusion**
* Through this project I was able to understand the factors that makes the detection of fake and real news easier.
* We have tested our data on 5 different models, viz., Naive Bayes, Logistic Regression, Decision Tree, Random Forest and SVM Classifier. We can see that all of them offers an accuracy score more than 95%, however, SVM or Support Vector Classifier offers utmost accuracy which is of 99.6%.