

FAKE NEWS DETECTION

Submitted by:

Rahul Kumar

**ACKNOWLEDGMENT**

It is indeed a great pleasure and matter of immense satisfaction for us to express our deep sense profound gratitude towards all the sources who have helped, inspired me in my project work.

I, would like to thanks **Flip Robo Technologies** which provided me this opportunity and was a constant source of inspiration for accomplishment of this project.

For the following project I used certain references which are mentioned below:

1. towardsdatascience.com
2. kaggle.com
3. stackoverflow.com
4. medium.com

**INTRODUCTION**

* Business Problem Framing

A fake news may contain sentences that are usually hoaxes and spread through social media or various other online platforms. Such news items may contain false and/or exaggerated claims, and may end up being vitalized by algorithms, and users may end up in a filter bubble. They may or may not contain spam words.

* Conceptual Background of the Domain Problem

The idea of fake news is not a novel concept. The idea has been in existence even before the emergence of the Internet as publishers used false and misleading information to further their interests. Following the advent of the web, more and more consumers began forsaking the traditional media channels used to disseminate information for online platforms. Not only does the latter alternative allow users to access a variety of publications in one sitting, but it is also more convenience and faster. The development, however, came with a redeﬁned concept of fake news as content publishers began using what has come to be commonly referred to as a clickbait.

* Review of Literature

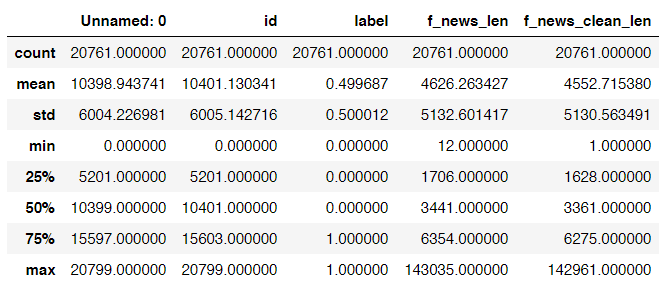
While we were working on this project we need to go through many sources, books and references. We are sharing some of the content which is relevant and useful in the lieu of this particular project.

* NumPy
* Pandas
* Sklearn
* Matplotlib
* Nltk
* Seaborn
* Wordcloud
* Tfidvectorizer
* Logistic Regression
* Decision Tree Classifier
* MultinomialNB
* Random Forest Classifier
* Pasive Aggressive Classifier
* Motivation for the Problem Undertaken

As we can see this is a highly motivated project. As this is the real time problem of fake news which is getting bigger, people are working hard to build good and real content and a single false news can ruin many things. Financial market is also based on news as we can observe based on new there are some fluctuation on Trading based on news.

**Analytical Problem Framing**

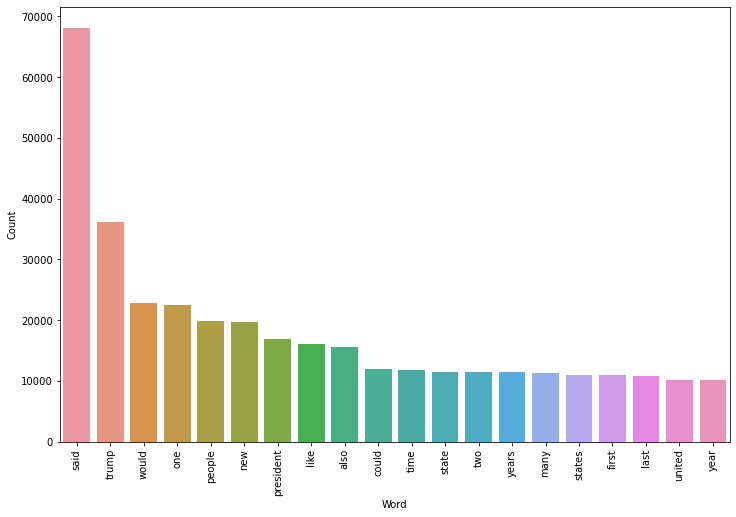
* Mathematical/ Analytical Modelling of the Problem



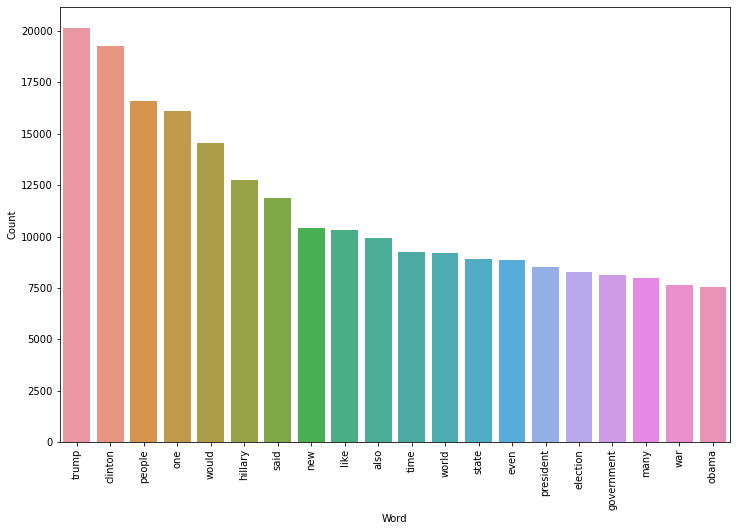
* Data Sources and their formats
* There are 6 columns in the dataset provided to you. The description of each of the column is given below:
* “id”: Unique id of each news article
* “headline”: It is the title of the news.
* “news”: It contains the full text of the news article
* “Unnamed:0”: It is a serial number
* “written\_by”: It represents the author of the news article
* “label”: It tells whether the news is fake (1) or not fake (0).
* Data Pre-processing Done



* Data Inputs- Logic- Output Relationships



Real News Word Count

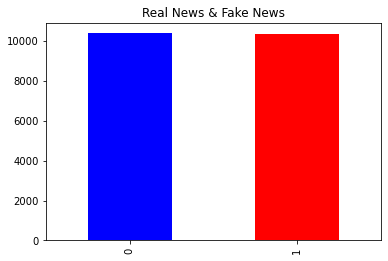


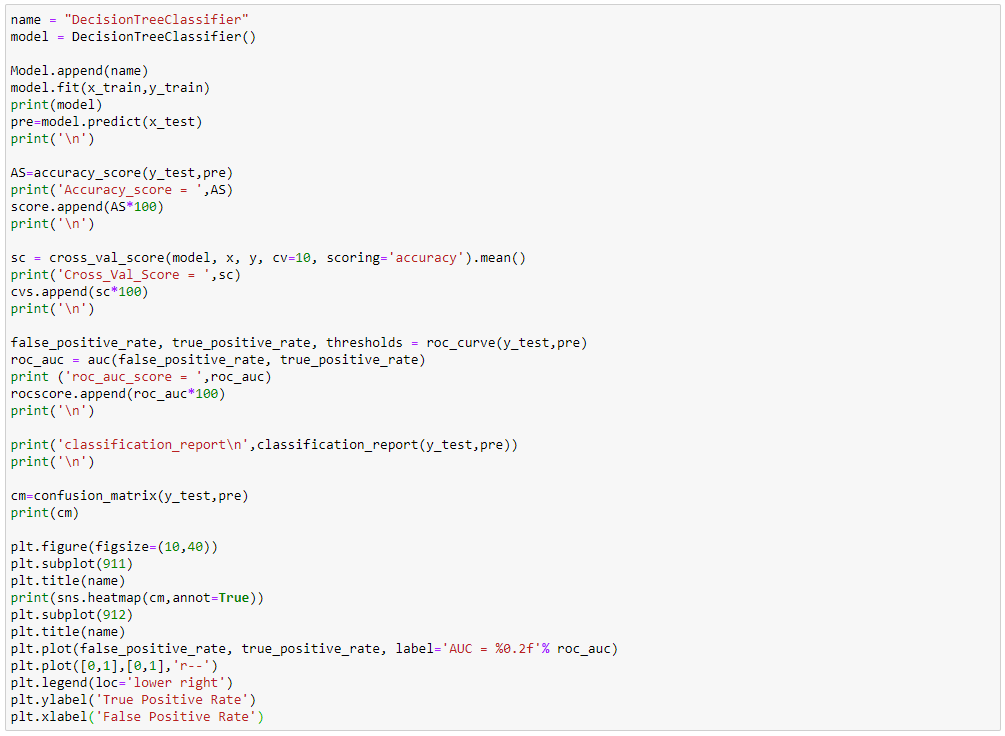
Fake News Word Count

* Hardware and Software Requirements and Tools Used
* **Hardware:** A good configuration Laptop
* **Software:** Python (Jupyter Notebook)
* **Library Used:** Mentioned above

**Model/s Development and Evaluation**

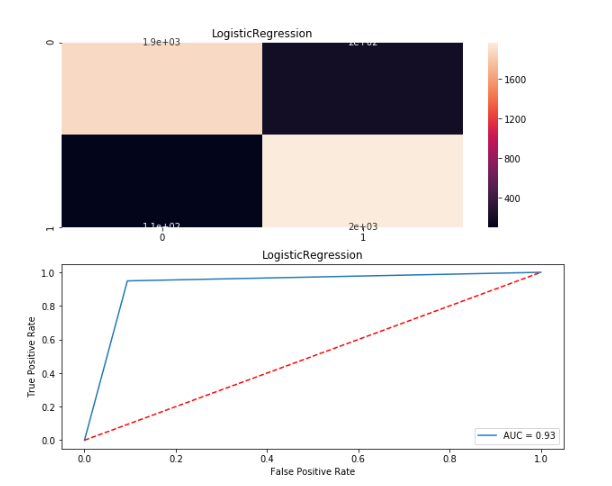
* Identification of possible problem-solving approaches (methods)

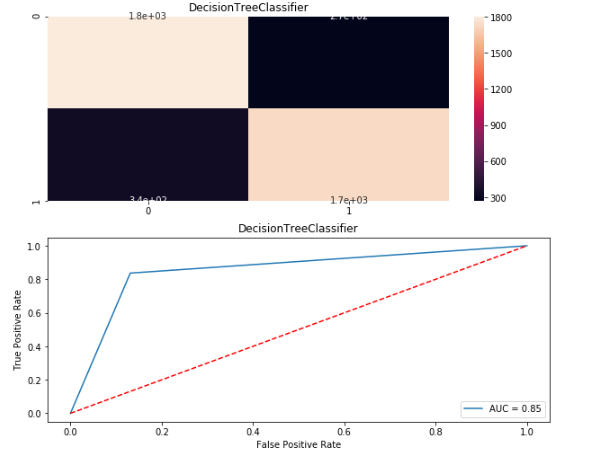


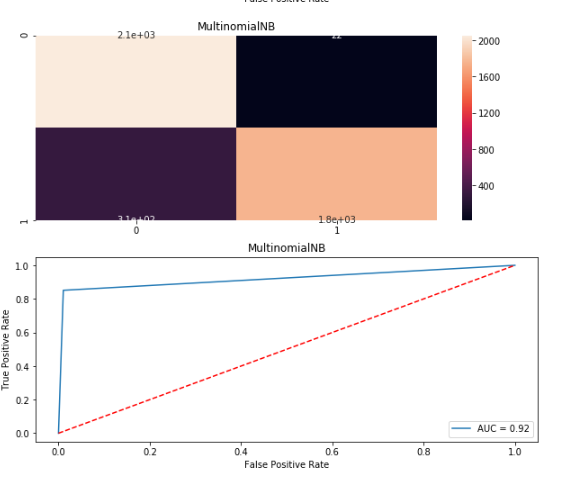
* Testing of Identified Approaches (Algorithms)
* Logistic Regression
* Decision Tree Classifier
* MultinomialNB
* Random Forest Classifier
* Pasive Aggressive Classifier
* Run and Evaluate selected models

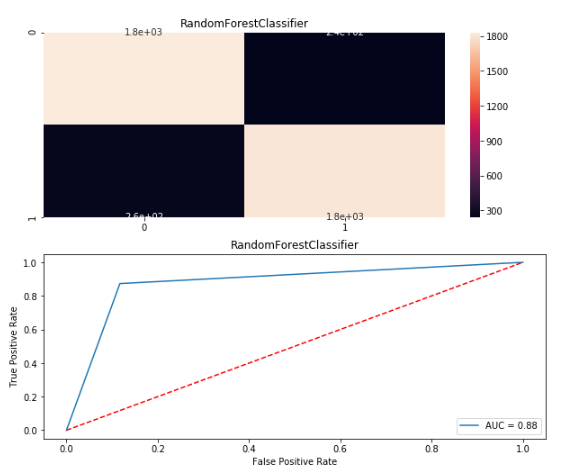


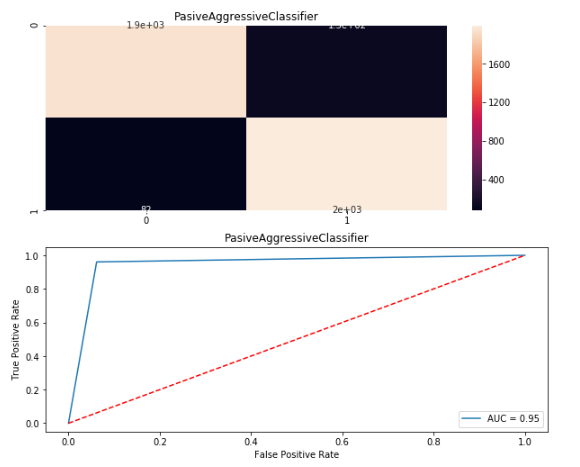
* Visualizations











* Interpretation of the Results

From visualizations and matrix we can see that the Passive Aggressive algorithms works best with 94% accuracy score, with cross val score of 95%, and Roc score of 94% and precision score of 96% and f1 score of 95% the max score which was achieved form the dataset provided.

**CONCLUSION**

* Key Findings and Conclusions of the Study

In the above dataset every features of the dataset plays an important role to understand the data also in visualization and applying models and algorithms.

* Learning Outcomes of the Study in respect of Data Science

Visualization is very helpful and it plays a vital role to understand the data into graphical form. So, that we can understand what the data is trying to say.

Data cleaning is also import part it helps me to remove null, nan or missing values by using mean, median, mode or by simply using 0.

I used five types of algorithm like LogisticRegression, DecisionTreeClassifier, MultinomialNB, RandomForestClassifier and PasiveAggressiveClassifier.

The problem I faced while on this dataset is it takes more than one hour for output. I overcome this situation with the help of towardsdatascience.com by running the code using small chunk of dataset to complete the project.