

***FAKE NEWS DETECTION PROJECT***

**Submitted by:**

**Shubham Sahu**

# ACKNOWLEDGMENT

I would like to thank Flip Robo Technologies for providing me with the opportunity to work on this project from which I have learned a lot. I am also grateful to Miss Khushboo Garg for her constant guidance and support.

Some of the reference sources are as follows:

* Internet
* Coding Ninjas
* Medium.com
* Analytics Vidhya
* StackOverflow

*TABLE OF CONTENTS*

[**ACKNOWLEDGMENT** 2](#_Toc75110441)

[**INTRODUCTION** 1](#_Toc75110442)

[BUSINESS PROBLEM FRAMING 1](#_Toc75110443)

[CONCEPTUAL BACKGROUND OF THE DOMAIN PROBLEM 1](#_Toc75110444)

[REVIEW OF LITERATURE 2](#_Toc75110445)

[MOTIVATION FOR THE PROBLEM UNDERTAKEN 2](#_Toc75110446)

[**ANALYTICAL** **PROBLEM** **FRAMING** 3](#_Toc75110447)

[MATHEMATICAL/ ANALYTICAL MODELING OF THE PROBLEM 3](#_Toc75110448)

[DATA SOURCES AND THEIR FORMATS 3](#_Toc75110449)

[DATA PREPROCESSING DONE 5](#_Toc75110450)

[DATA INPUTS- LOGIC- OUTPUT RELATIONSHIPS 9](#_Toc75110451)

[HARDWARE AND SOFTWARE REQUIREMENTS AND TOOLS USED 9](#_Toc75110452)

[**MODEL/S DEVELOPMENT AND EVALUATION** 14](#_Toc75110453)

[IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES (METHODS) 14](#_Toc75110454)

[TESTING OF IDENTIFIED APPROACHES (ALGORITHMS) 15](#_Toc75110455)

[RUN AND EVALUATE SELECTED MODELS 16](#_Toc75110456)

[KEY METRICS FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION 24](#_Toc75110457)

[**CONCLUSION** 25](#_Toc75110458)

[KEY FINDINGS AND CONCLUSIONS OF THE STUDY 25](#_Toc75110459)

[LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE 25](#_Toc75110460)

[LIMITATIONS OF THIS WORK AND SCOPE FOR FUTURE WORK 25](#_Toc75110461)

# INTRODUCTION

## BUSINESS PROBLEM FRAMING

News media has become a channel to pass on the information of what’s happening in the world to the people living. Often people perceive whatever conveyed in the news to be true. There were circumstances where even the news channels acknowledged that their news is not true as they wrote. But some news has a significant impact not only on the people or government but also on the economy. One news can shift the curves up and down depending on the emotions of people and political situation.

It is important to identify the fake news from the real true news. The problem has been taken over and resolved with the help of Natural Language Processing tools which help us identify fake or true news based on historical data. The news is now in safe hands!

## CONCEPTUAL BACKGROUND OF THE DOMAIN PROBLEM

The authenticity of Information has become a longstanding issue affecting businesses and society, both for printed and digital media. On social networks, the reach and effects of information spread occur at such a fast pace and so amplified that distorted, inaccurate, or false information acquires a tremendous potential to cause real-world impacts, within minutes, for millions of users. Recently, several public concerns about this problem and some approaches to mitigate the problem were expressed.

The sensationalism of not-so-accurate eye-catching and intriguing headlines aimed at retaining the attention of audiences to sell information has persisted all throughout the history of all kinds of information broadcast. On social networking websites, the reach and effects of information spread are however significantly amplified and occur at such a fast pace, that distorted, inaccurate, or false information acquires a tremendous potential to cause real impacts, within minutes, for millions of users.

## REVIEW OF LITERATURE

Fake news is not a new concept. Before the era of digital technology, it was spread through mainly yellow journalism with a focus on sensational news such as crime, gossip, disasters and satirical news. With the widespread dissemination of information via digital media platforms, it is of utmost importance for individuals and societies to be able to judge the credibility of it. Fake news is not a recent concept, but it is a commonly occurring phenomenon in current times. The consequence of fake news can range from being merely annoying to influencing and misleading societies or even nations. A variety of approaches exist to identify fake news

## MOTIVATION FOR THE PROBLEM UNDERTAKEN

The widespread problem of fake news is very difficult to tackle in today’s digital world where there are thousands of information sharing platforms through which fake news or misinformation may propagate. It has become a greater issue because of the advancements in AI which brings along artificial bots that may be used to create and spread fake news. The situation is dire because many people believe anything they read on the internet and the ones who are amateur or are new to the digital technology may be easily fooled. A similar problem is fraud that may happen due to spam or malicious emails and messages. So, it is compelling enough to acknowledge this problem take on this challenge to control the rates of crime, political unrest, grief, and thwart the attempts of spreading fake news. Text, or natural language, is one form that is difficult to process simply because of various linguistic features and styles like sarcasm, metaphors, etc. Moreover, there are thousands of spoken languages and every language has its grammar, script and syntax. Natural language processing is a branch of artificial intelligence and it encompasses techniques that can utilize text, create models and produce predictions. This work aims to create a system or model that can use the data of past news reports and predict the chances of a news report being fake or not.

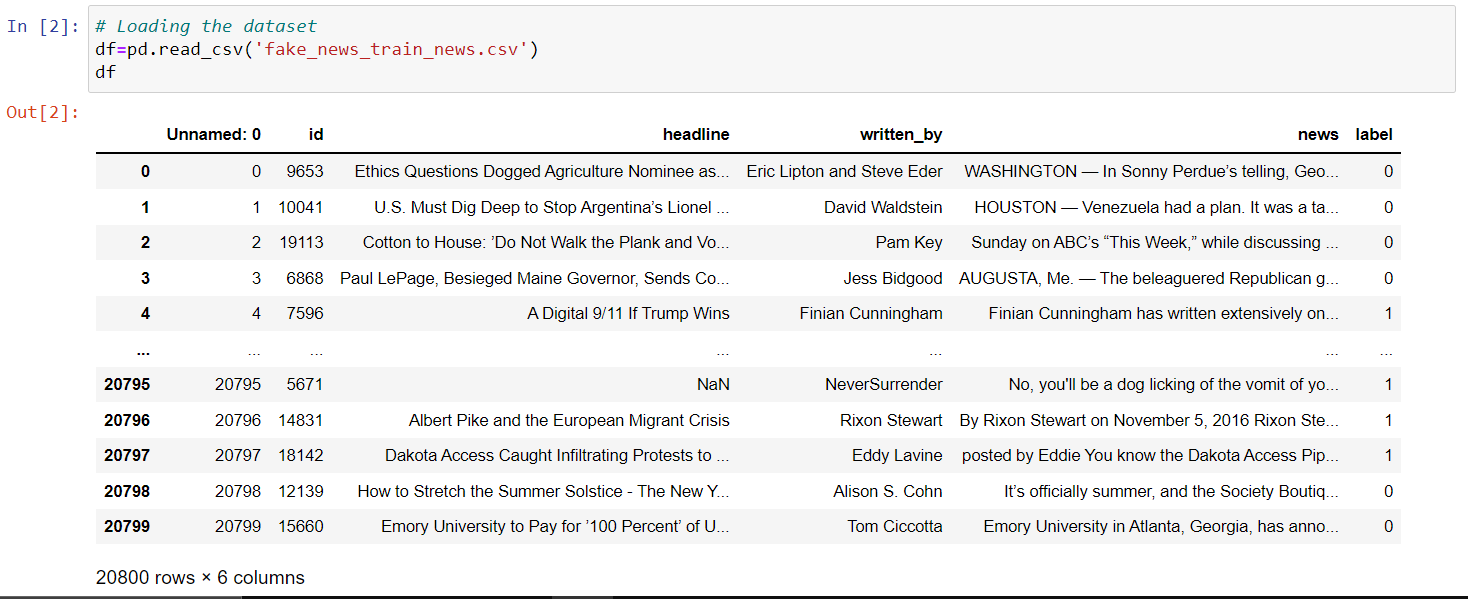
# ANALYTICAL PROBLEM FRAMING

## MATHEMATICAL/ ANALYTICAL MODELING OF THE PROBLEM

* The dataset provided here has a shape of (20800, 6). Which means it has 20800 rows and 6 columns?
* The target or the dependent variable named “Label” has two distinct values 0 and 1. Where 0 represents the news that is not fake or authentic while 1 represents the category of fake news. As the target column „Label‟ is giving binary outputs and all the independent variables has text so it is clear that it is a supervised machine learning problem where we can use, we can use the techniques of NLP and classification-based algorithms of Machine learning.
* Here we will use NLP techniques like word tokenization, lemmatization and tfidf vectorizer then those processed data will be used to create the best model using various classification based supervised machine learning algorithms like Logistic Regression, Multinomial NB, Random Forest Classifier etc
* The dataset contains null value.
* Train test is the best way to get the solution of these kinds of problems as that is the easiest and the efficient way to solve this problem.

## DATA SOURCES AND THEIR FORMATS

* The data is provided to us from our client database. The sample data is in .csv format
* The sample data for reference is shown below.



**Dataset description**

There are 6 columns in the dataset provided:

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).

**Identification of possible problem-solving approaches (methods)**

We have used the following process for problem-solving:

1. Data Preprocessing

2. Building a word dictionary

3. Feature extraction

4. Training classifiers

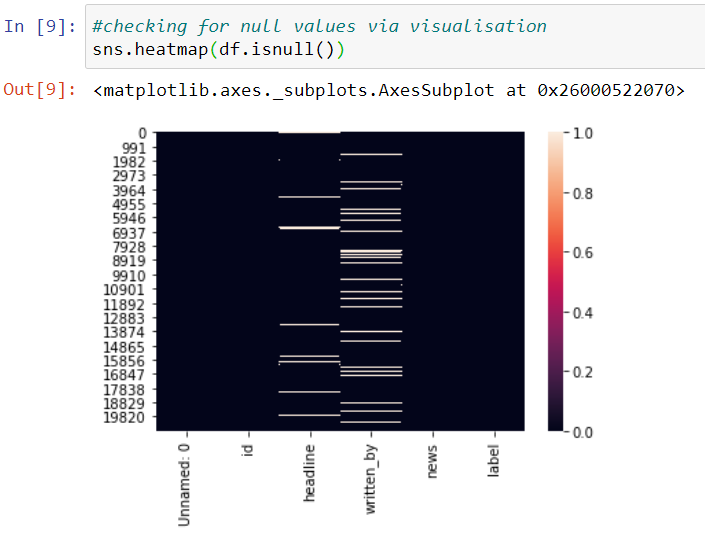
5. Testing

6. Performance evaluation using multiple metrics

## DATA PREPROCESSING DONE

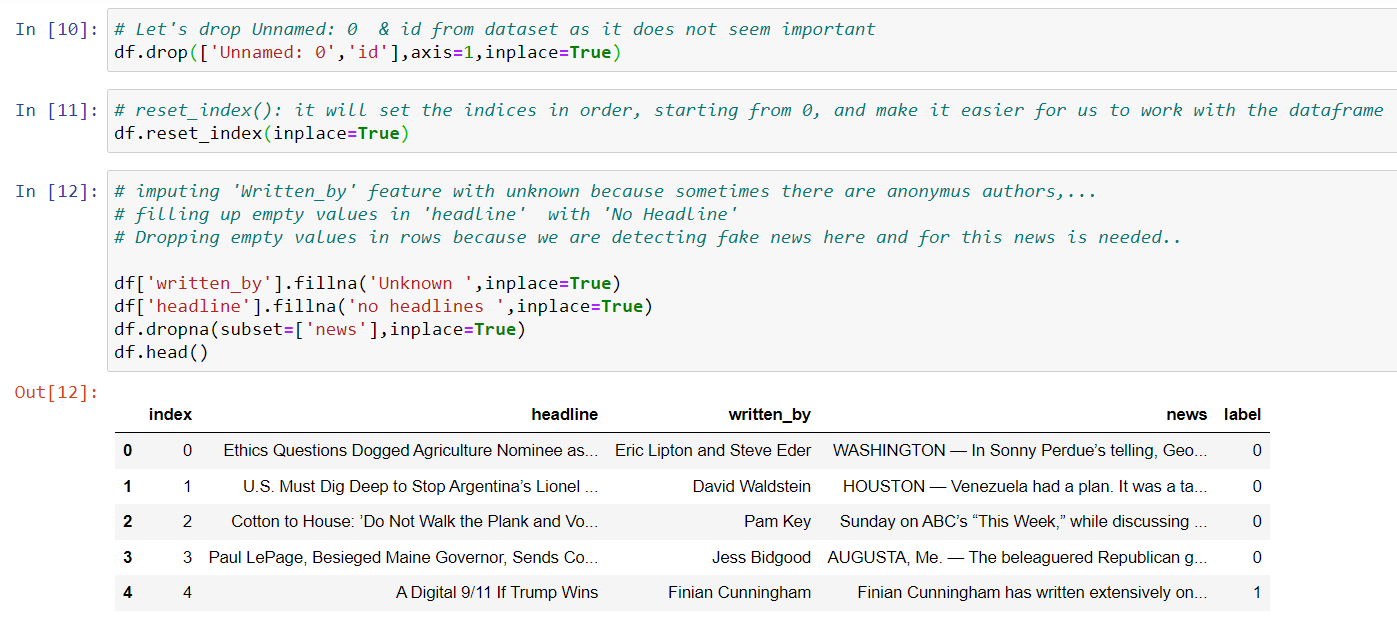
Data usually comes from a variety of source & is often inconsistent, inaccurate. Data preprocessing helps to enhance the quality of data and make it ready for the various ML model. We have applied various methods for data preprocessing methods in this project.

* First, we check shape by using (df. shape)
* Then checked datatype of various features & found that all features are of int type except headline, written\_by, news which are of object datatype
* Checking for null values in each column



It clearly shows that null values are present in the dataset, which needs to be removed.

**Treating null values**



**Checking distribution of fake and real news**

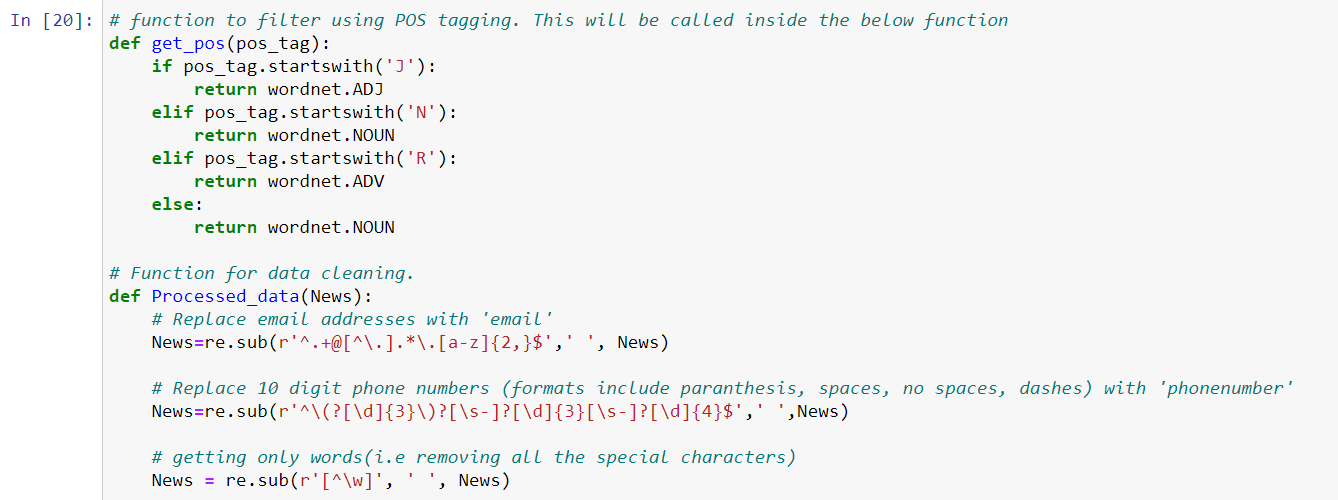


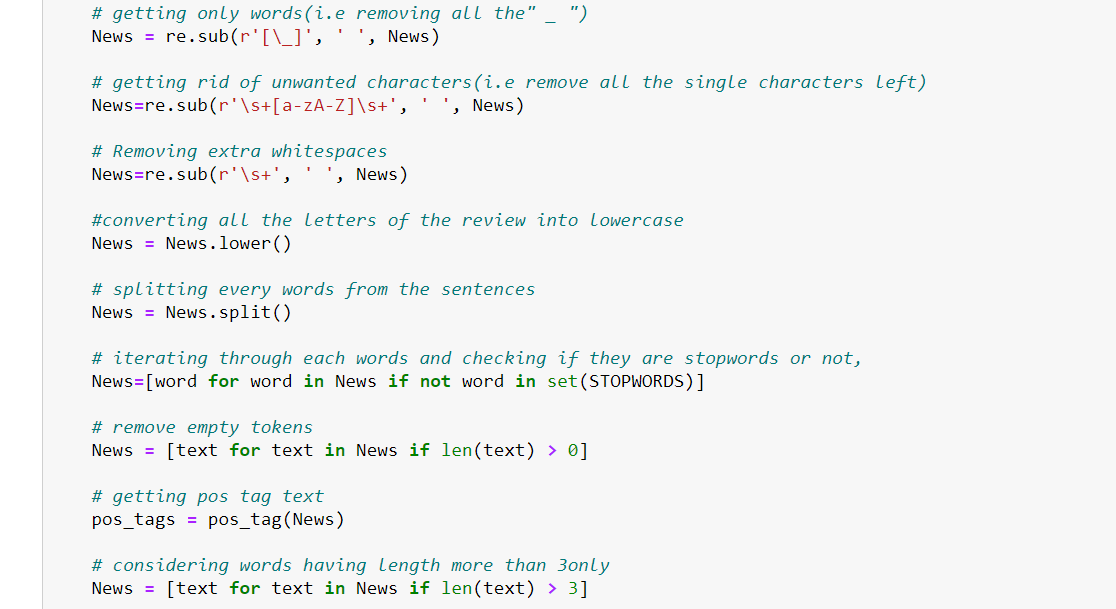


We see that both news is equally distributed .ie dataset is balanced which is good as it will help our model to classify more accurately, so we should expect a good accuracy score.

Cleaning the raw data-It involves the deletion of words or special characters that do not add meaning to the text. Important cleaning steps are as follows:

1. Lowering case
2. Handling of special characters
3. Removal of stopwords
4. Handling of hyperlinks
5. Removing leading and trailing white space
6. Replacing URLs with web address
7. Converted words to the most suitable base form by using lemmatization



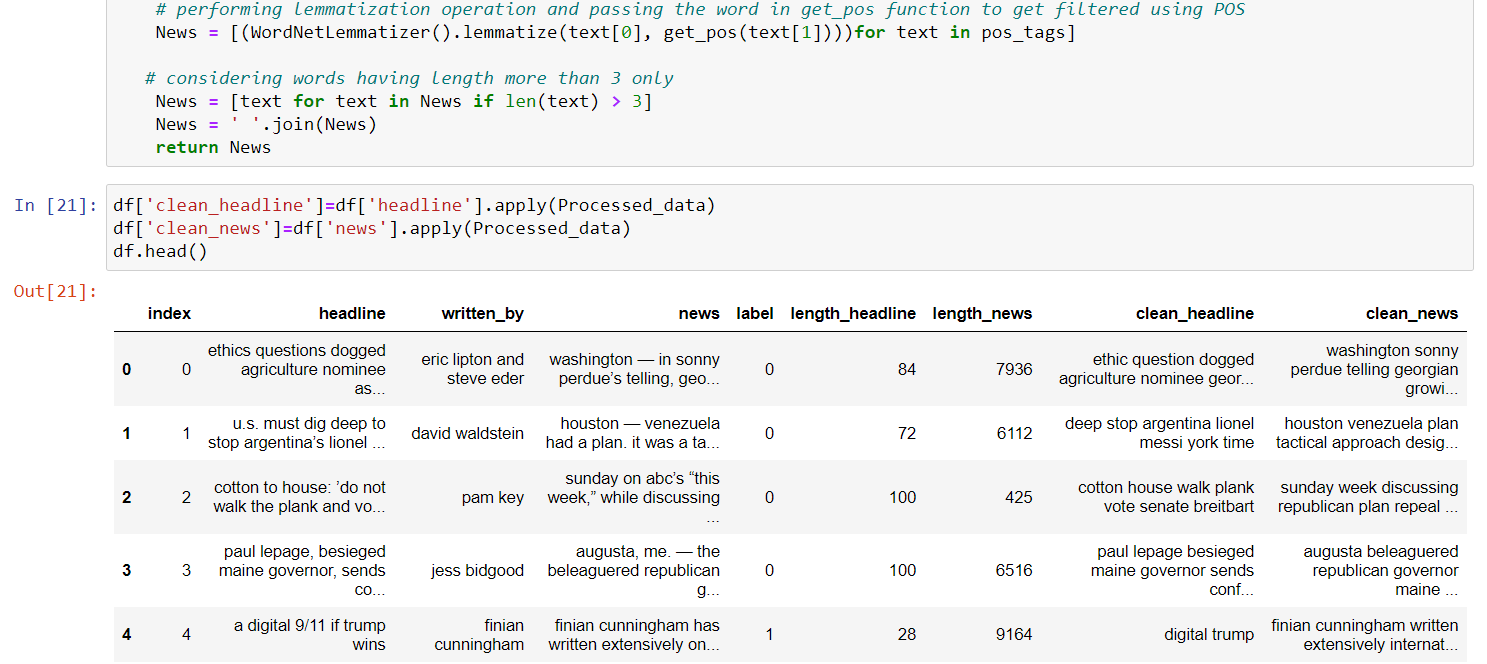


For Data pre-processing we did some data cleaning, where we used WordNet lemmatizer to clean the words and removed special characters using Regexp Tokenizer and filter the words by removing stop words and then used lemmatizers and joined and return the filtered words.

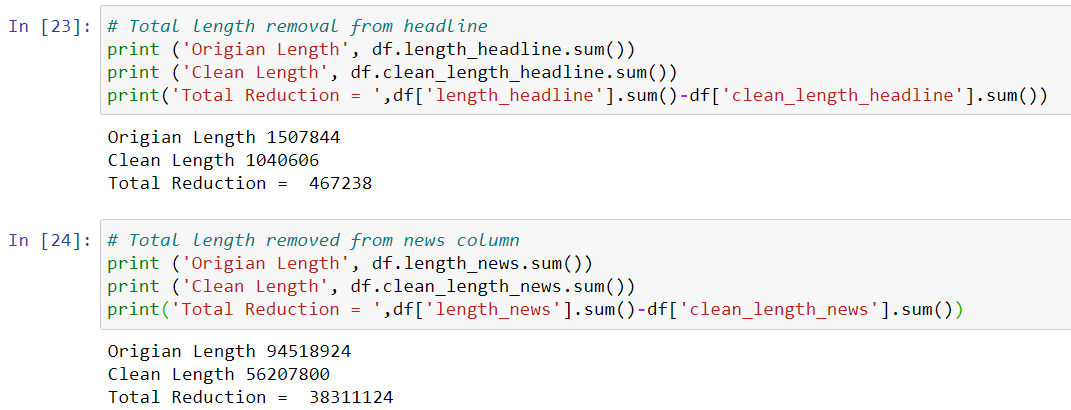
Used TFIDF vectorizer to convert those text into vectors, and split the data and into test and train and trained various Machine learning algorithms.

**Adding additional attribute:**

To compare the length of headline & news before preprocessing and after preprocessing an addition column was added:







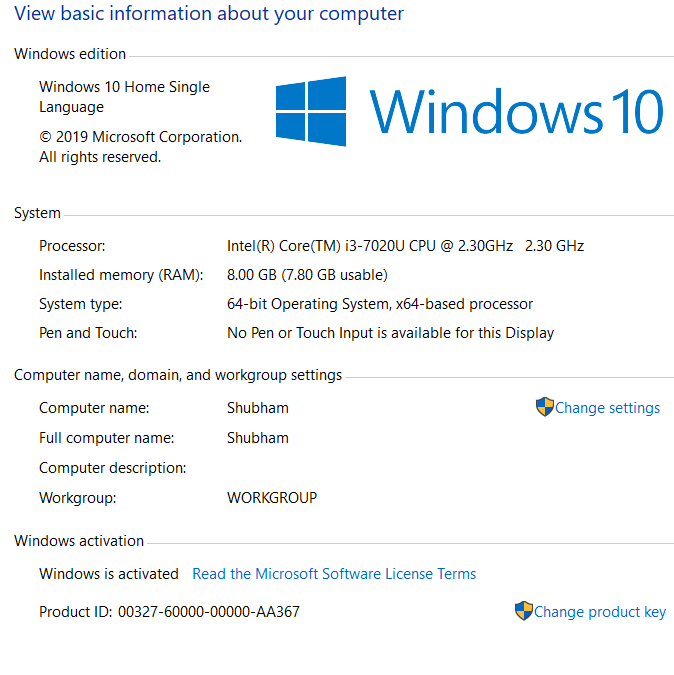
After executing all these steps it was found that all the words & special characters were removed from the dataset which was of no use and consuming memory

## DATA INPUTS- LOGIC- OUTPUT RELATIONSHIPS

For this data’s input and output logic, we will analyse words frequency for each label, so that we can get the most frequent words that were used in different features.

## HARDWARE AND SOFTWARE REQUIREMENTS AND TOOLS USED

***HARDWARE:***



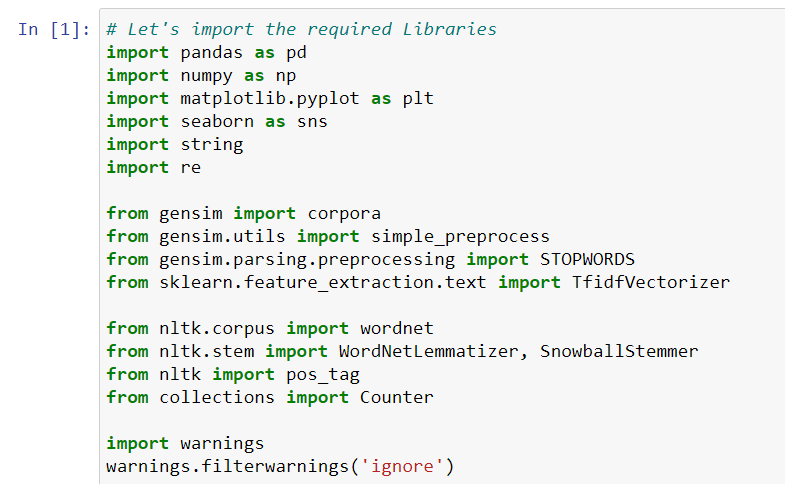
***SOFTWARE:***

Jupyter Notebook (Anaconda 3) – Python 3.7.6

Microsoft Excel 2010

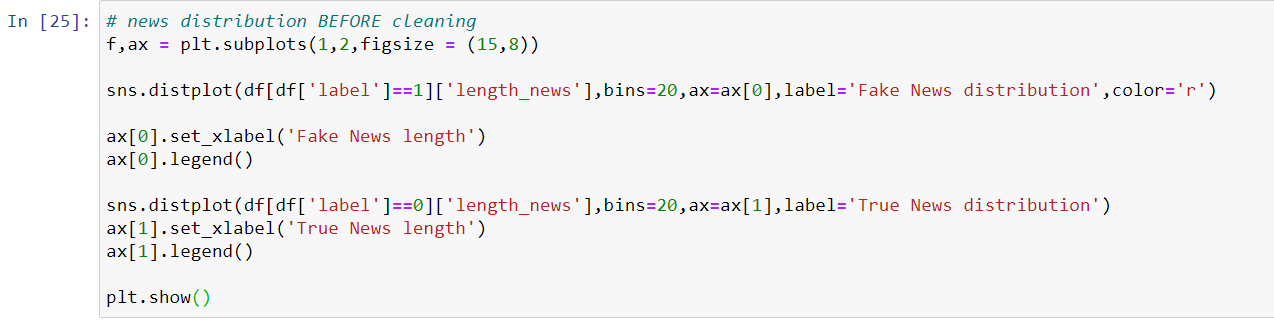
***LIBRARIES:***

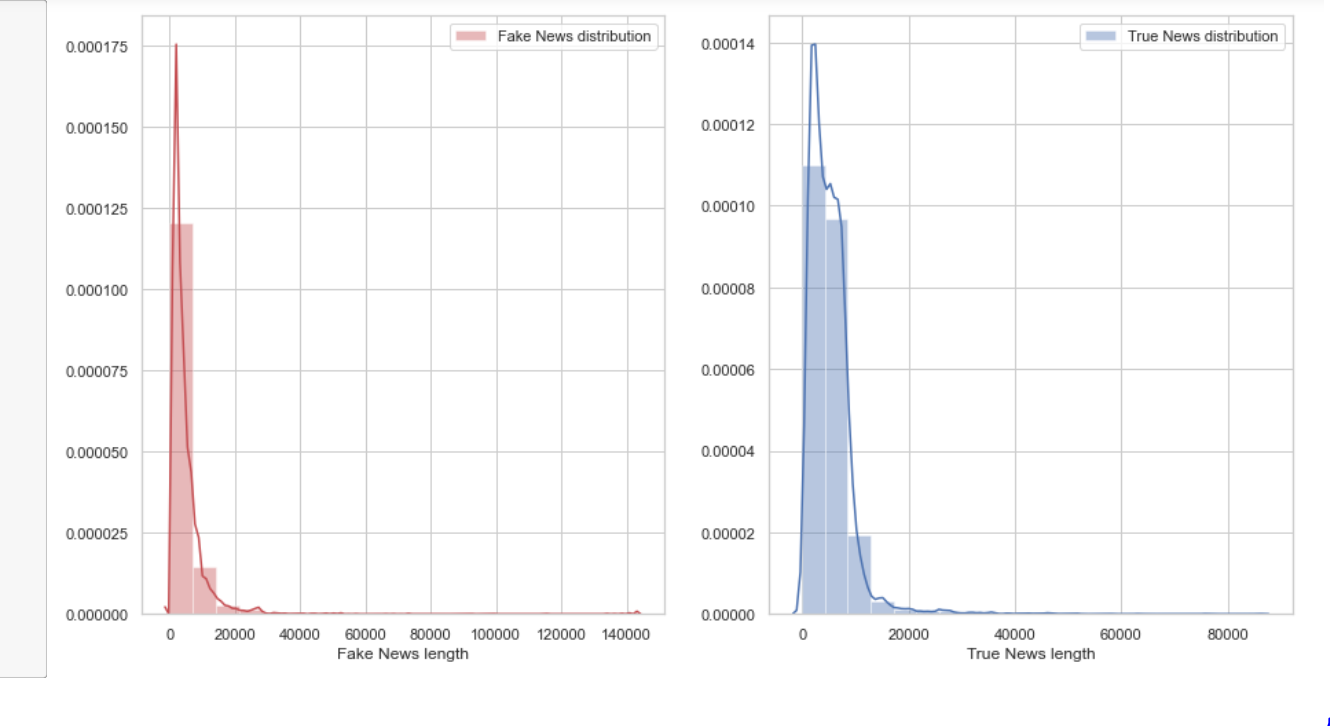
* Pandas: To read the Data file in form of data.
* Matplotlib: This library is typically used to plot the figures for better visualisation of data.
* Seaborn: A advanced version of Matplotlib
* Scikit Learn: This is the most important library for Machine Learning since it contains various Machine Learning Algorithms which are used in this project. Scikit Learn also contains Preprocessing library which is used in data preprocessing. Apart from this, it contains a very useful joblib library for serialization purpose using which the final model has been saved in this project.
* NLTK: Natural language took kit is one of the most used libraries for building NLP projects.



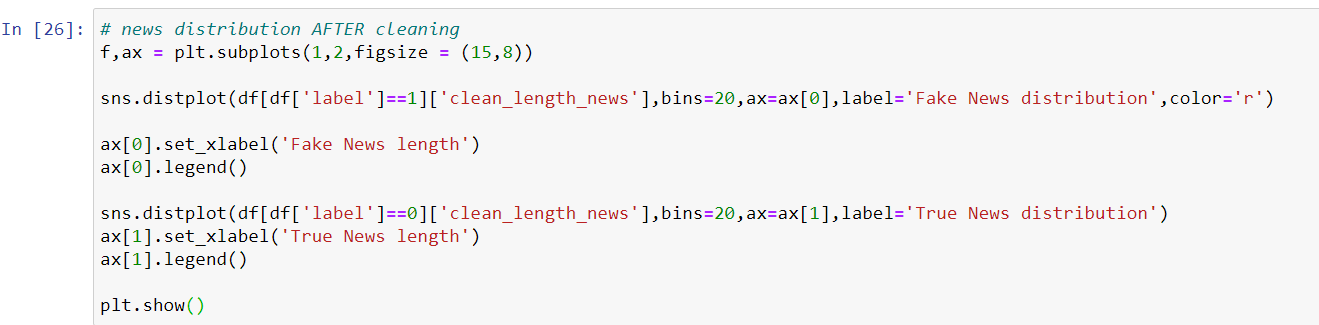
**Then we have plotted a graph to show the distribution of word count before cleaning and after cleaning**

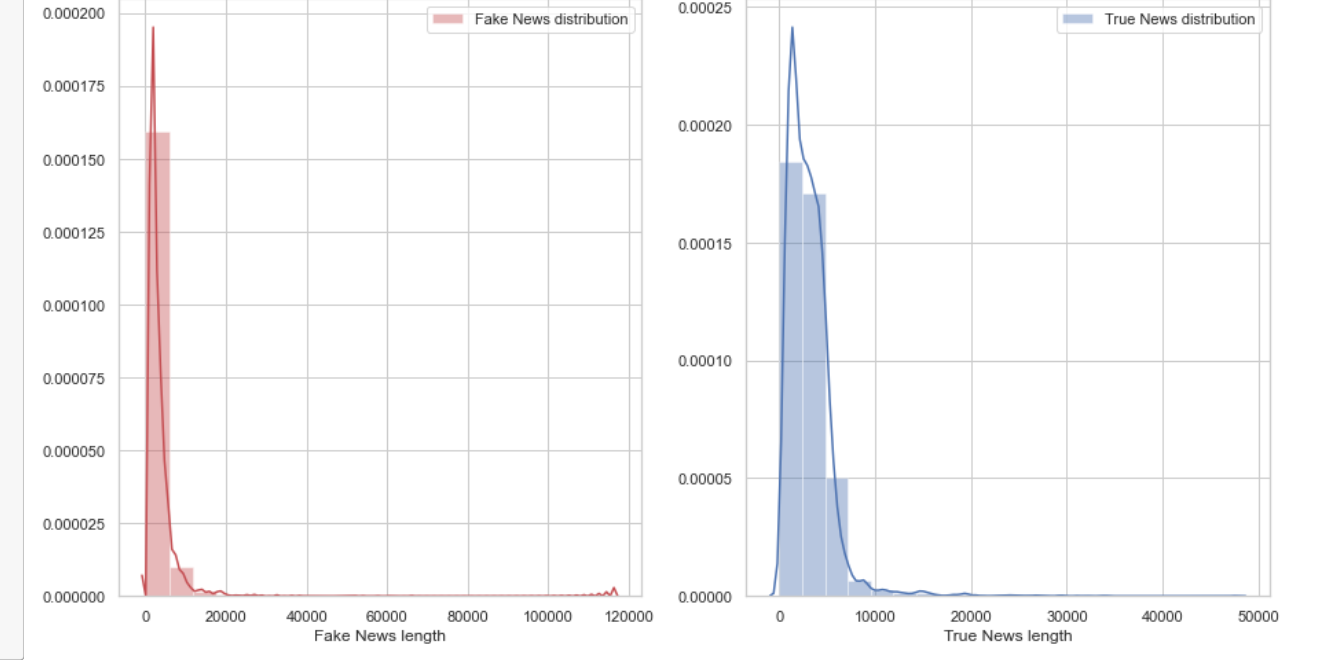
**Before cleaning:**





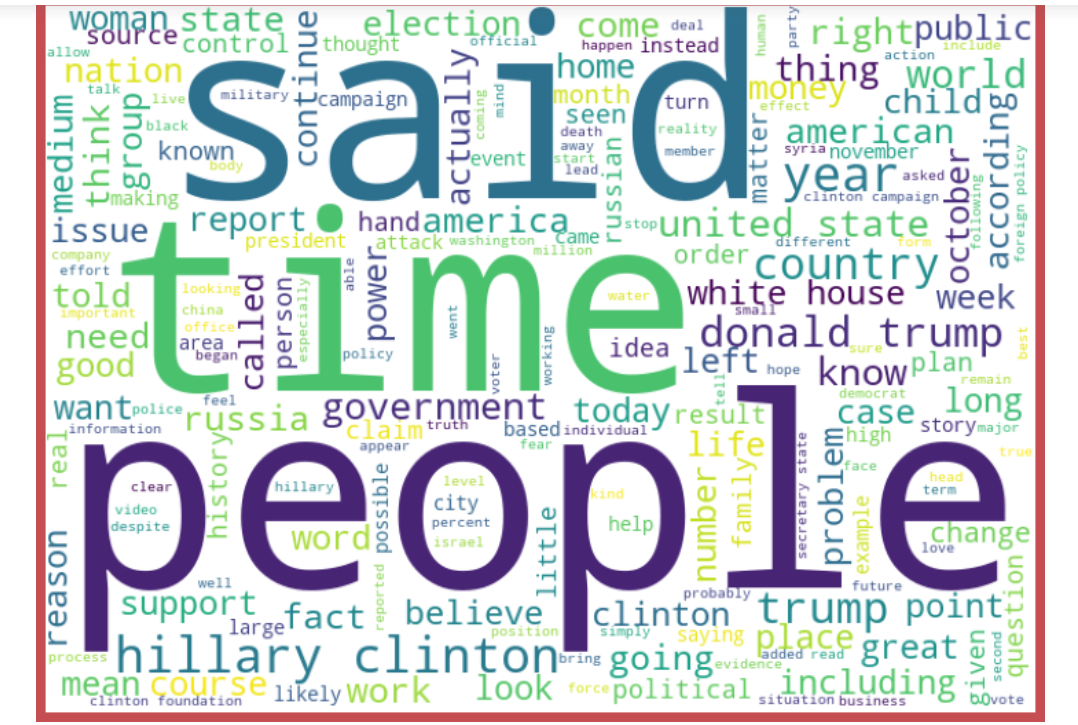
**After cleaning**



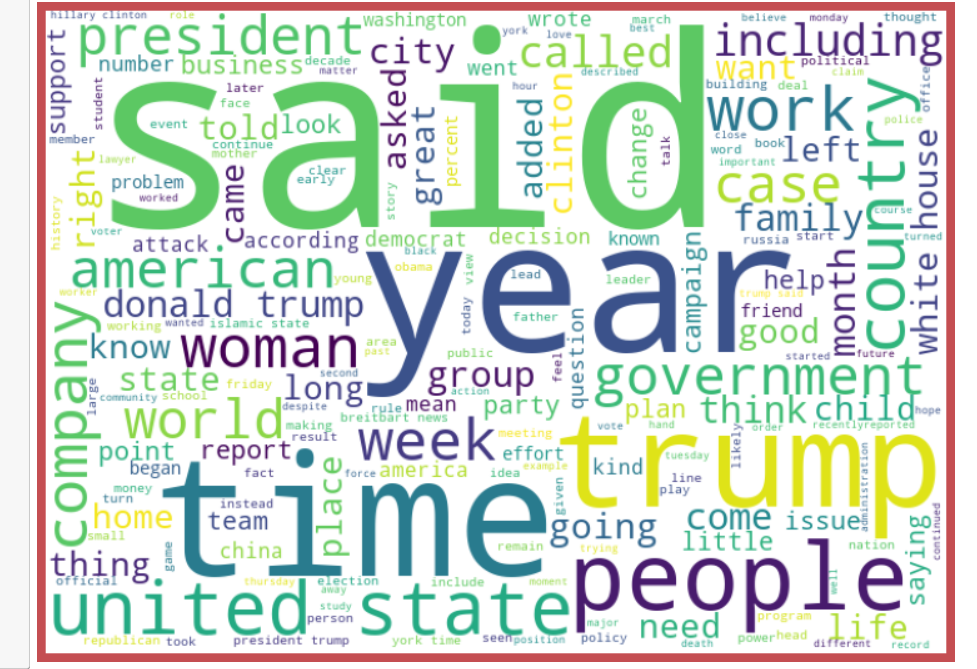


To get a better view of words contained in news. A word dictionary (word cloud ) was made showing the words highly occurred in fake and real news for both headline and news column.

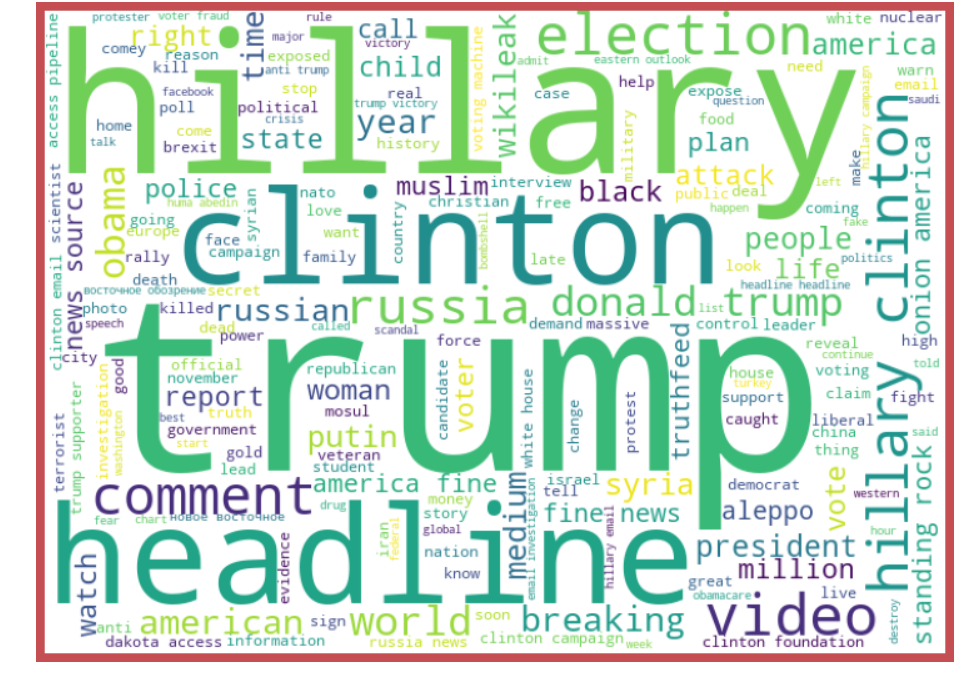


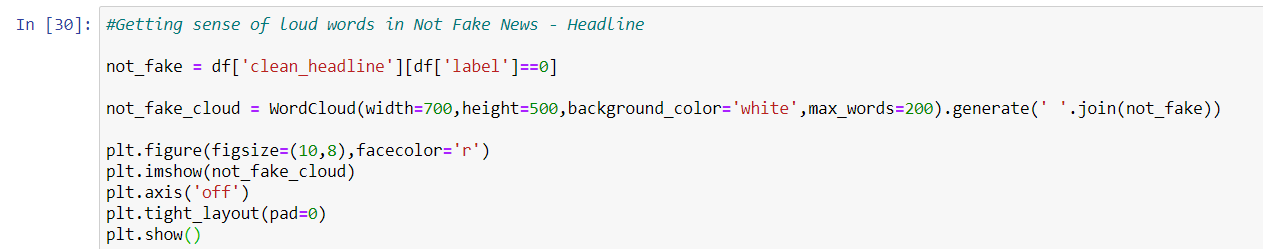


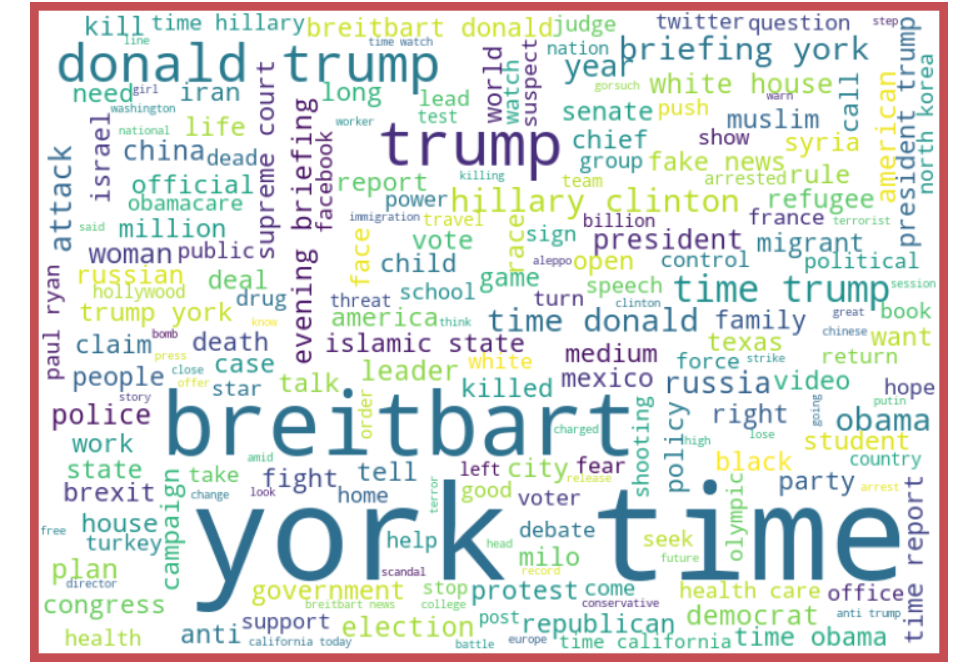












From the above we can see that the most frequent words on both labels and we can observe the words which are leading to fake news are trump, Clinton, prison, November, etc. and words which are leading to real news are said, agriculture, police, questions etc., so we can see that above dataset extensively deals with news around US presidential elections between Trump and Clinton.

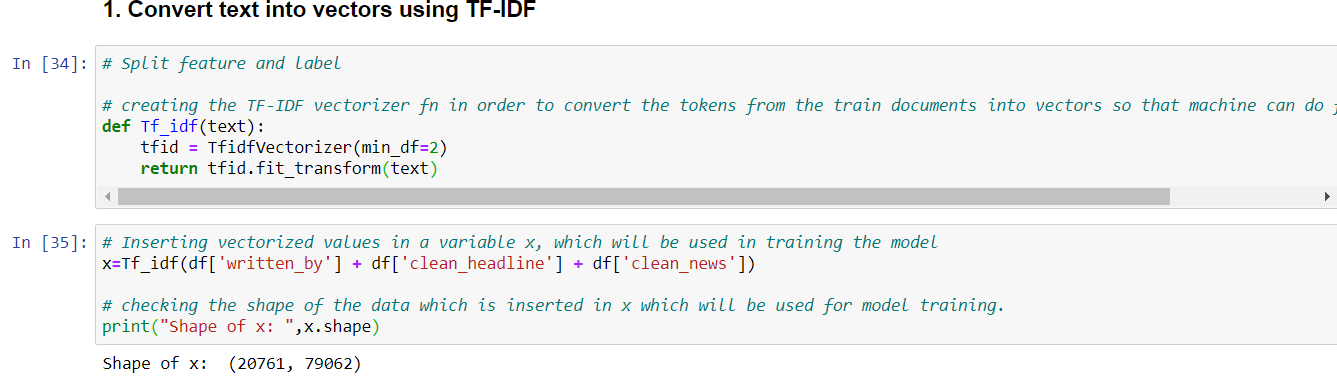
# MODEL/S DEVELOPMENT AND EVALUATION

## IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES (METHODS)

Understanding the problem is the first crucial steps in solving any problem. From the given dataset it can be concluded that it is a binary classification problem. Therefore I run my preprocessed data on 6 classification algorithm.

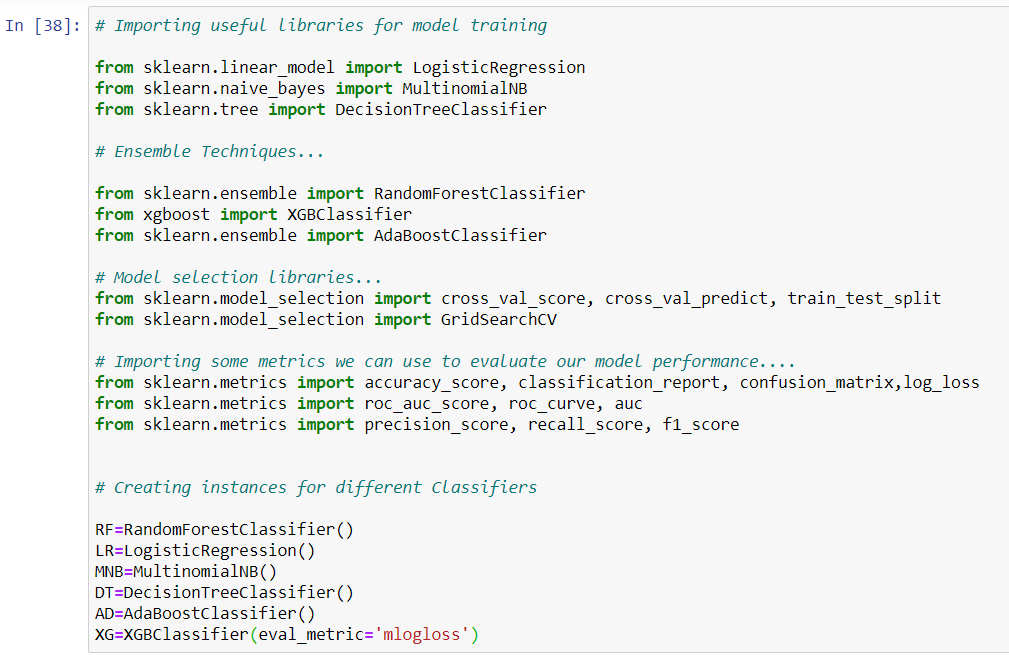
**Training Classifier:**

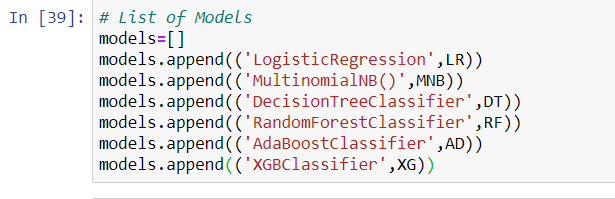
We converted all the text into vectors, using TF-IDF. Then we have split features and label.



## TESTING OF IDENTIFIED APPROACHES (ALGORITHMS)

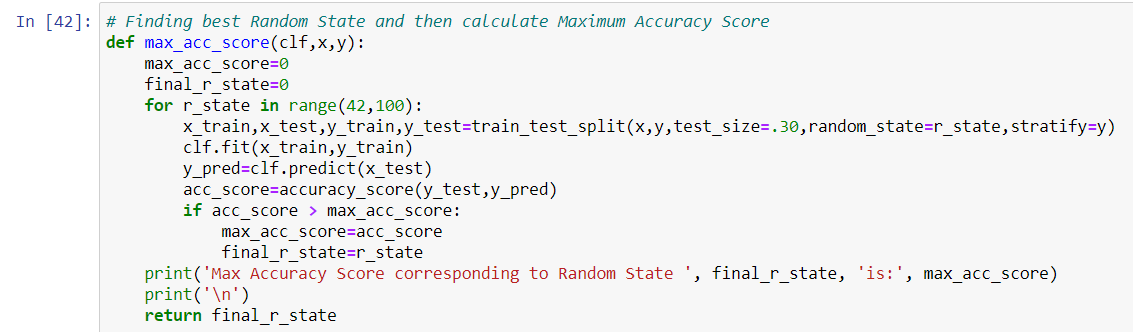
The algorithms we used for the training and testing are as follows:-

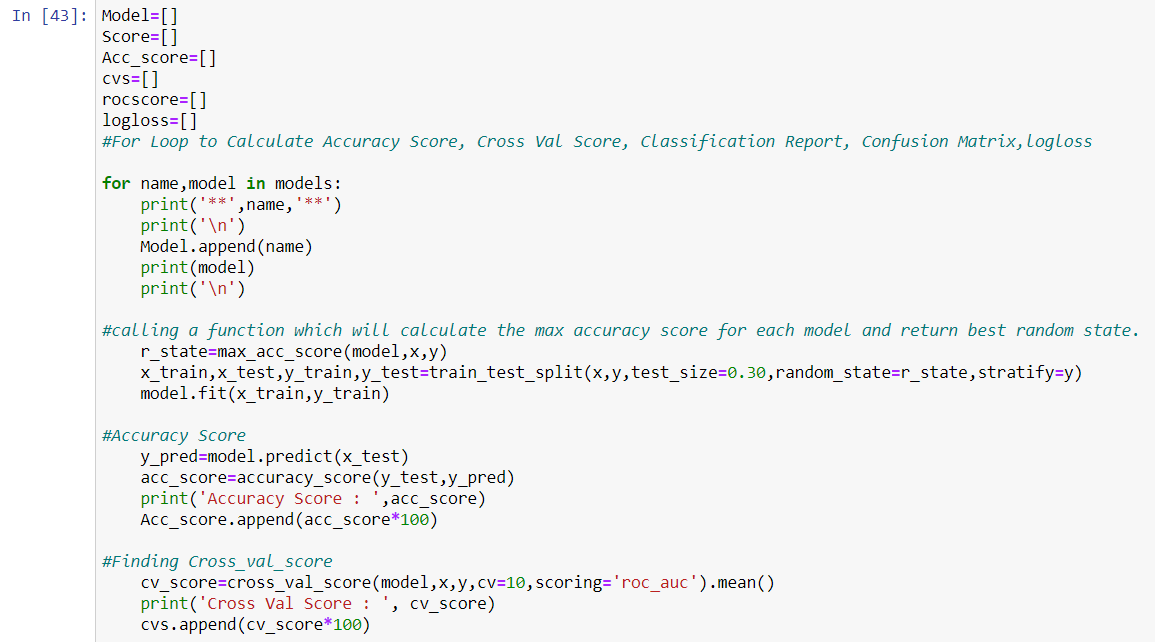


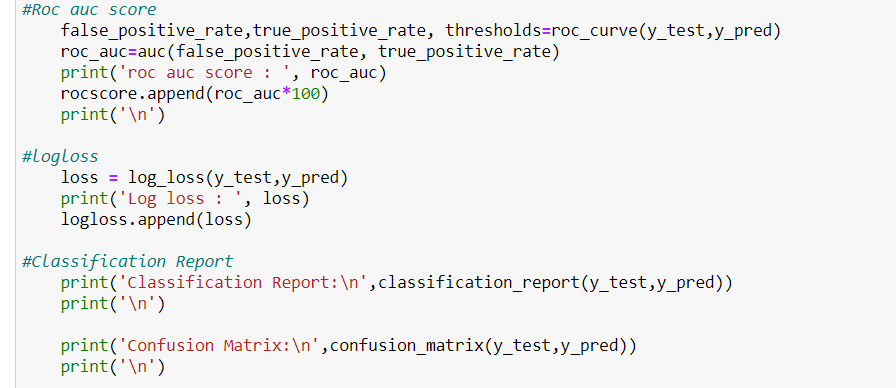


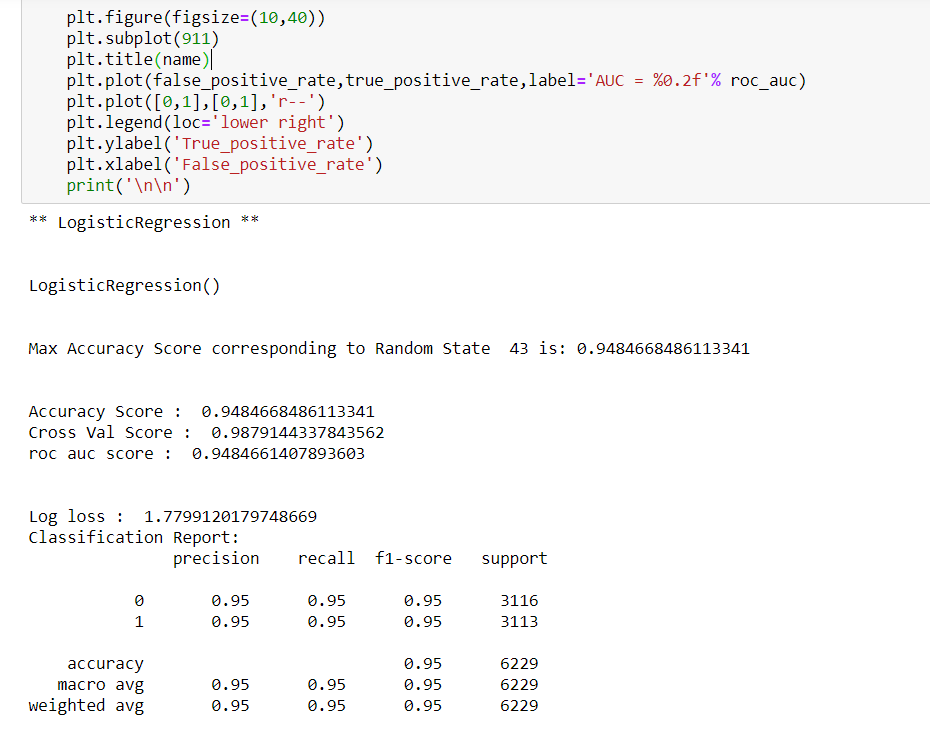
## RUN AND EVALUATE SELECTED MODELS

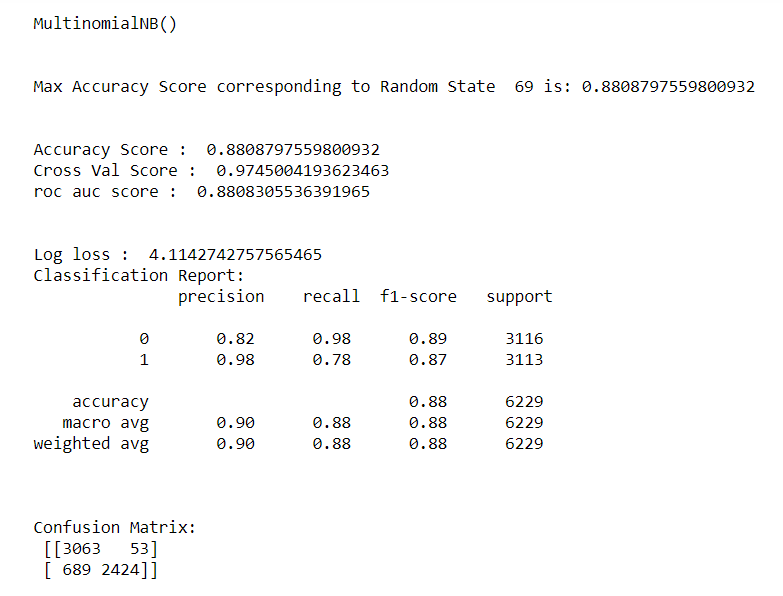
In my approach, I have first prepared a method that gives all necessary classification metrics of an algorithm like classification metrics, auc\_roc score, confusion matrix, log\_loss.

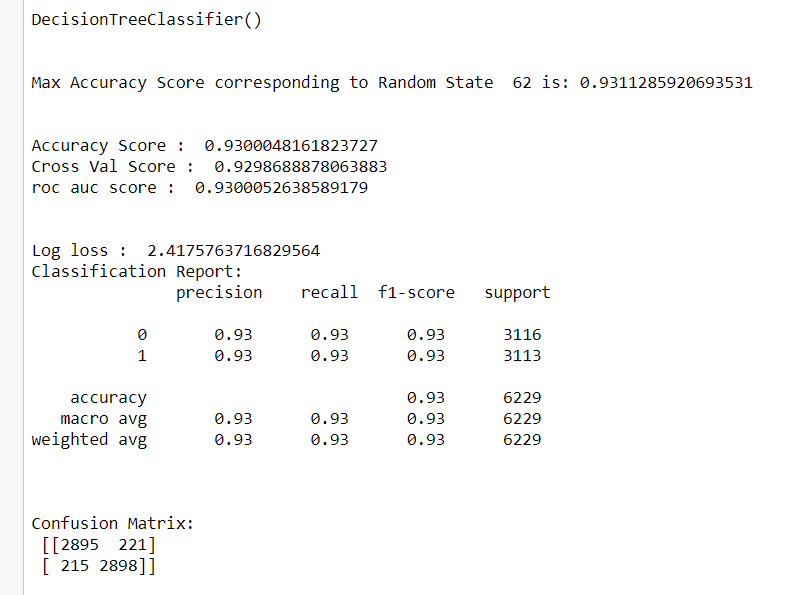


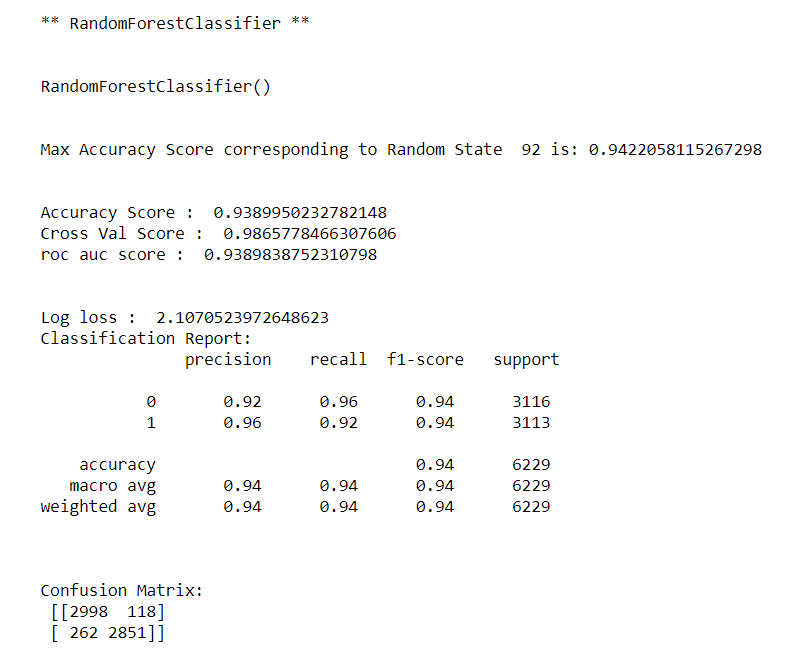


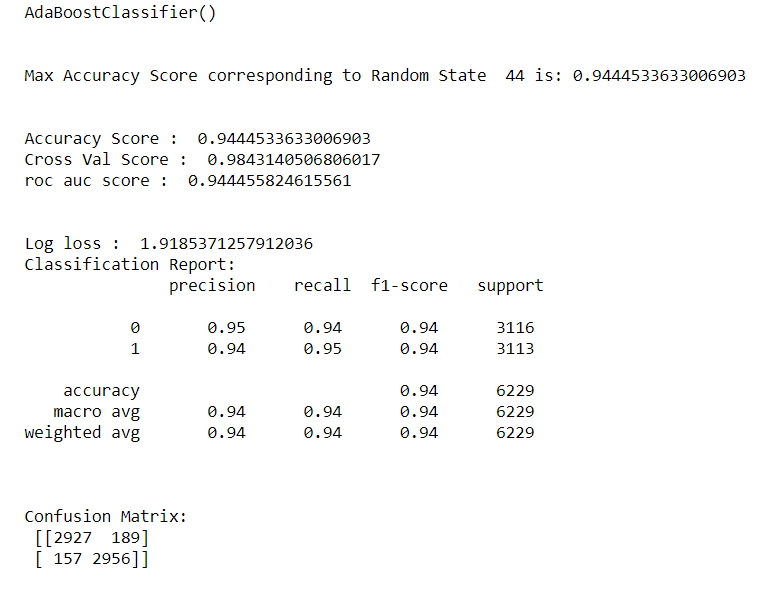


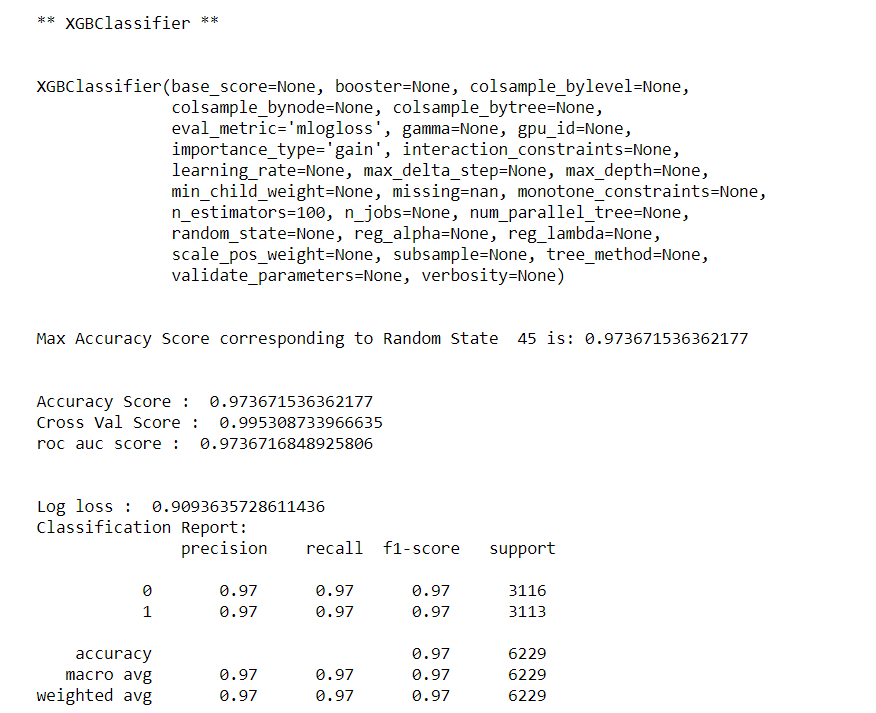


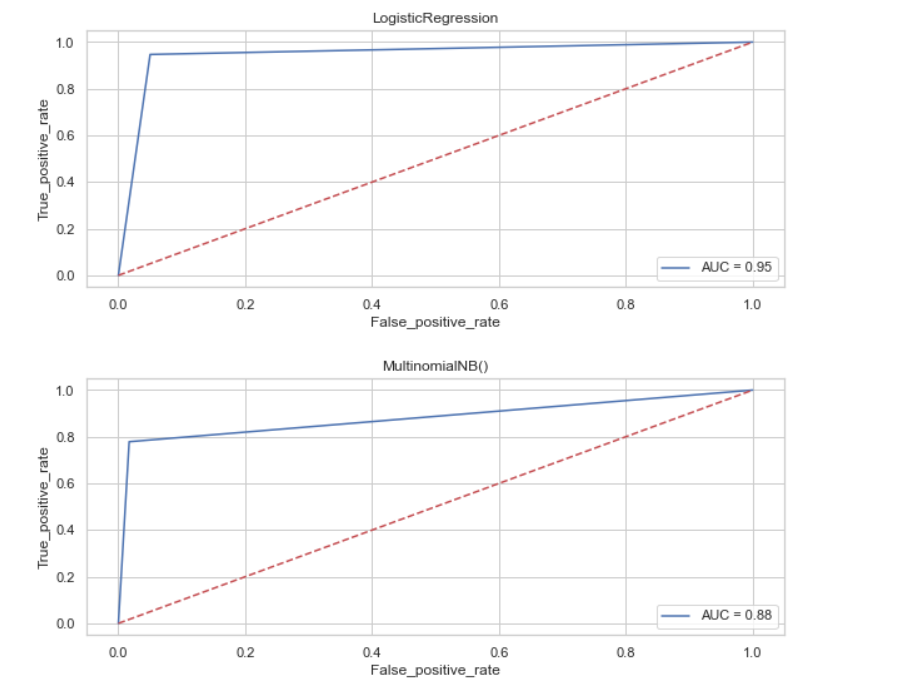


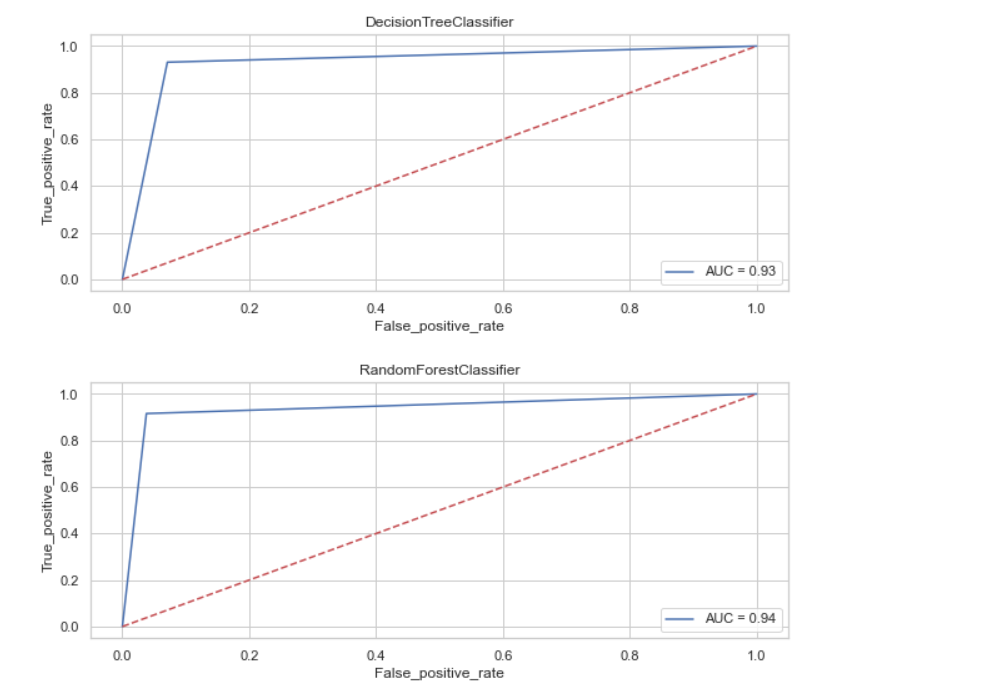


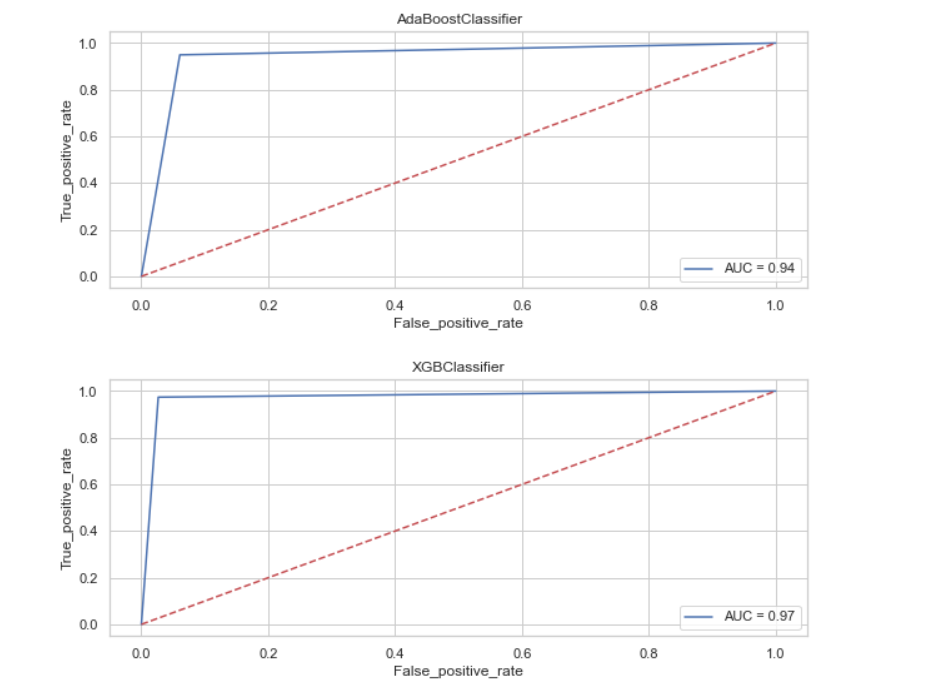


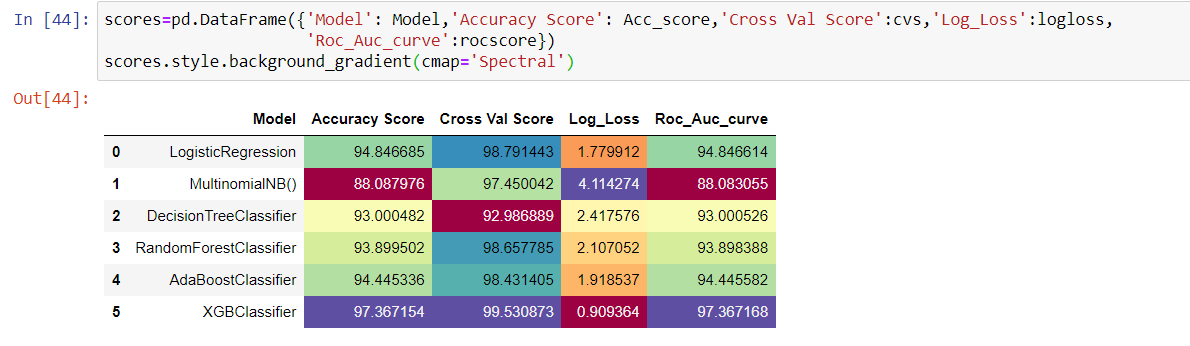






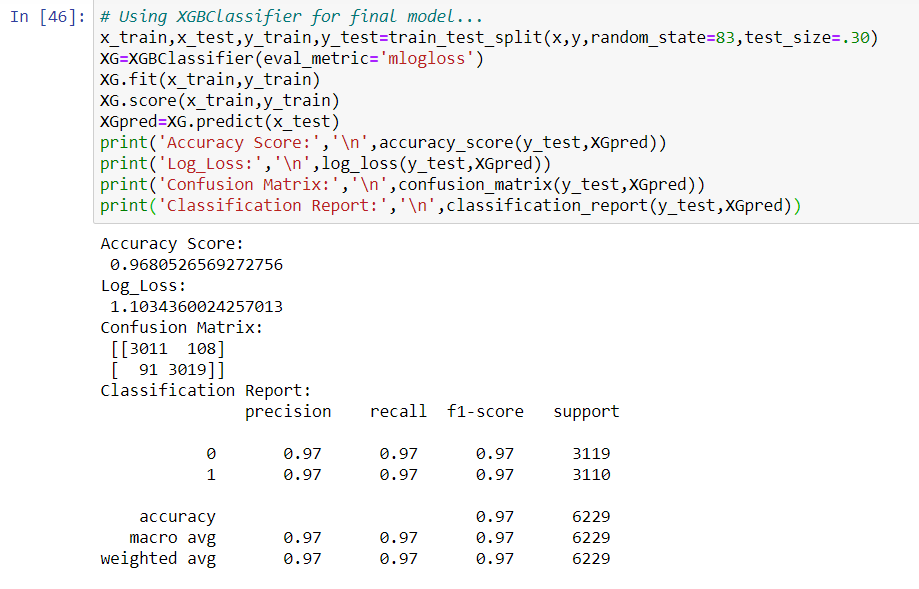


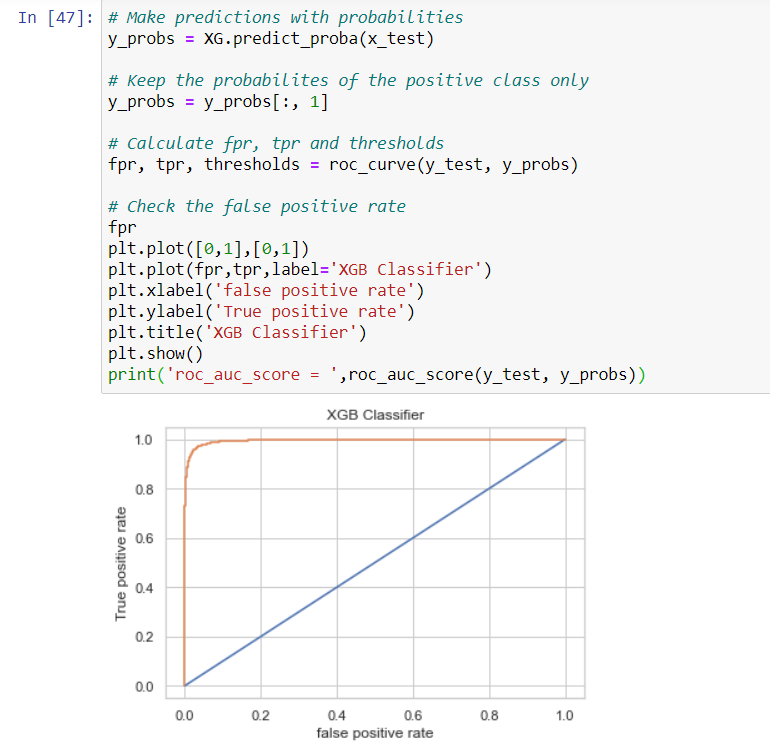


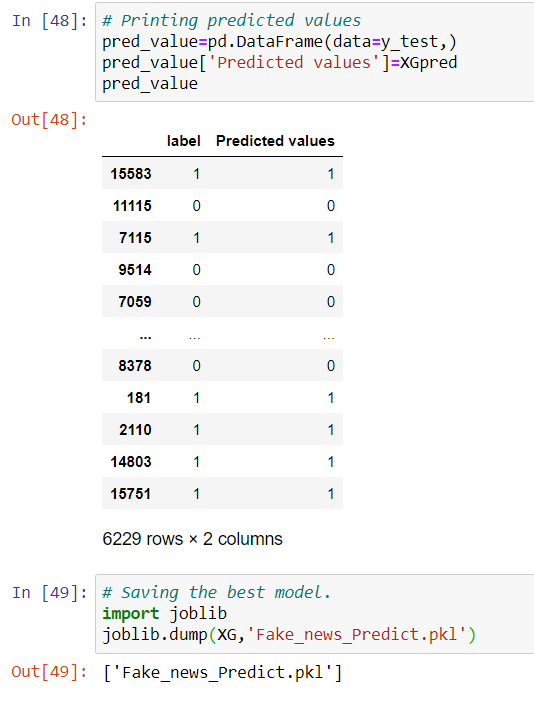


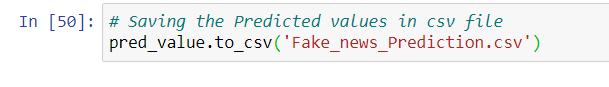
We choose the XGBoost Classifier model as the final one, as it gives the highest accuracy score & also log\_loss value is minimum which indicates the better prediction

**FINAL MODEL**









## KEY METRICS FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION

* When it comes to the evaluation of a data science model’s performance, sometimes accuracy may not be the best indicator.
* Some problems that we are solving in real life might have a very imbalanced class and using accuracy might not give us enough confidence to understand the algorithm’s performance.
* In the fake news problem that we are trying to solve, the data is balanced. so accuracy score nearly tells the right predictions. So the problem of overfitting in this problem is nearly not to occur. So here, we are using an accuracy score to find a better model.

# CONCLUSION

## KEY FINDINGS AND CONCLUSIONS OF THE STUDY

From the whole evaluation, we can see that the maximum number of words in fake news were regarding Trump, and Clinton and we can interpret that it was due to election campaign which was held during the US presidential election and we know these adverse effects of the voters which were influenced by the fake news and most of the real news had said, trump and president, and fake news which was cleared by trump’s campaign, but can hardly see any clarity or real news from the side of Clinton, and due to which the impact we already saw on election results and regarding the election advertisement and news Facebook’s CEO Mark Zuckerberg also got extensively question by congress.

## LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE

It is possible to classify news content into the required categories of authentic and fake news however there will be always a bias to this kind of classification which depends on the behavioural pattern of the listener. However, using this kind of project awareness can be created to know what is fake and authentic.

## LIMITATIONS OF THIS WORK AND SCOPE FOR FUTURE WORK

Machine Learning Algorithms like XGBoost, Adaboost and Randomforest Classifier took an enormous amount of time to build the model. Using Hyper-parameter tuning for XGB would have resulted in some more accuracy.