Fake News Detection Model

Submitted by:

**Ishant Guleria**

**ACKNOWLEDGMENT**

I would like to express my special thanks of gratitude to all the teachers who have taught me NLP and Machine Learning due to knowledge they had provided me I am able to complete this project.

**INTRODUCTION**

* **Business Problem Framing**

In Today’s World information can easily be accessible from anywhere ,Today individual can access the happenings of various events around the world in the comfort of his/her own home. It has resulted in the inaccuracy and irrelevancy in updating information by people which is commonly known as fake news which can cause real-world impacts, within minutes, for millions of users. Since a large proportion of the population uses social media for updating themselves with news, delivering accurate and altruistic information to them is of utmost importance. Due to the increasing number of users in social media, news can be quickly published by anybody, and its credibility stands compromised, As fake news is written to mislead readers, it makes it a difficult task to detect based on the content of the news only. The news content is diverse in terms of styles, the subject in which it is written, it becomes essential to bring an efficient system for its detection. Fake news detection has recently garnered much attention from researchers and developers . This work proposes to detect fake news using various methods using NLP and machine learning.

* **Review of Literature**

If we look at some scholar work shows the issue that the fake news has been major concerned amongst scholar from various background. For instance, some authors have observed that fake news is no longer a preserve of the marketing and public relations departments. Instead there is a increasing risk of IT security, therefore, IT department is premised on the idea that it would help avert the various risks associated with the problem. So, if we good deeply into it we could find that the hackers use click bait with the help of fake news and make some professional of the organization downloads their malicious exploits in their system or leak sensitive information, albeit in an indirect manner. The user may, for instance, be tricked into believing that they are helping to disseminate the news further when, in the actual sense, they are providing the perpetrators with access to their emails, and we can also see that the fake news are worked extensively as they are using videos with original massage and uses their facial structure to replace the massage with false massage they want us to believe, these fake news issues is bigger day by day and we need to implement more our research and extensive knowledge to solve the problem.

* **Motivation for the Problem Undertaken**

As in today’s digital world in which we are connected to web 24\*7 we came across different news , many time these news are fake which can cause lot of damage to the society , country and world even some time it has also caused riots that why fake news is very serious issue . So it is very necessary to build a model that can detect the news whether it is fake or not which will help to keep only the authentic news in the world.

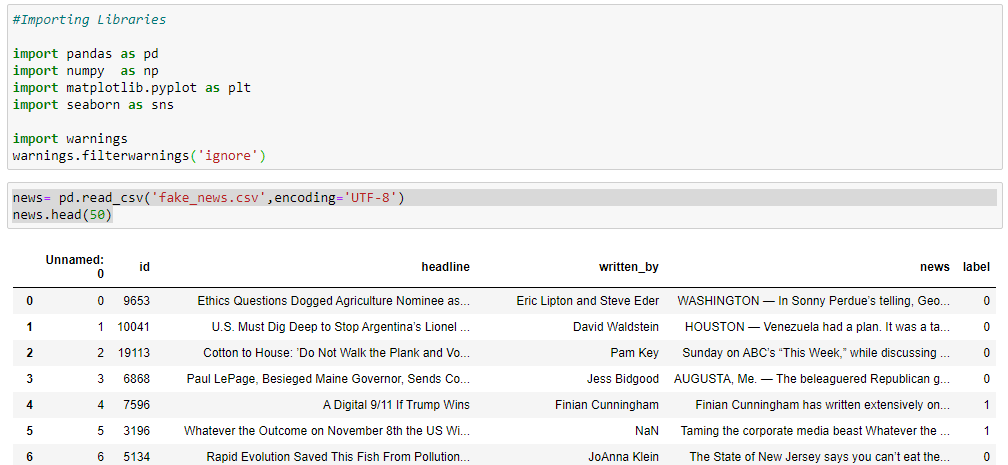
**Analytical Problem Framing**

* **Data Sources and their formats**

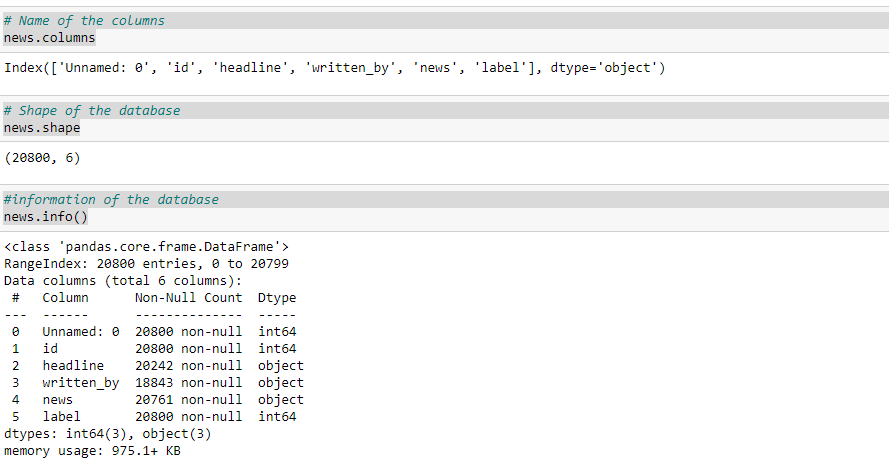
In this project we are given a CSV file of name fake\_news.csv . we will use the fake\_news.csv data to build a model to predict whether a news is fake or not fake.

* **Mathematical/ Analytical Modeling of the Problem**

First of all we will load necessary libraries and then will load our fake\_news.csv file .

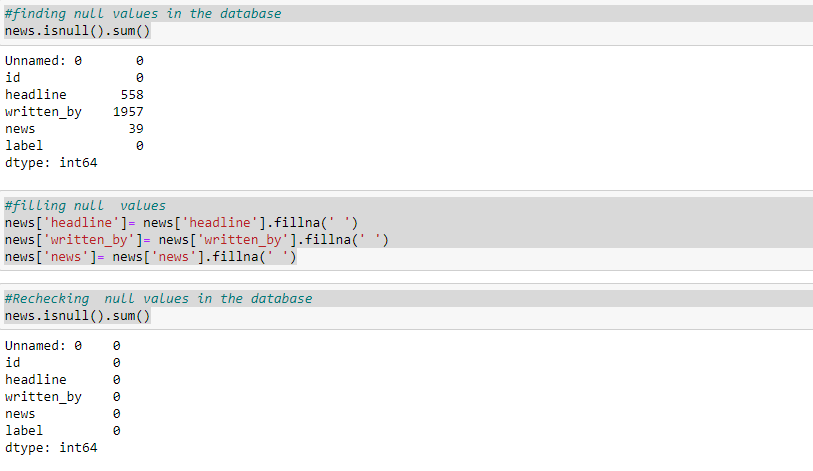


Now will check the name of the columns , shape of the database and information about the database.



* **Data Preprocessing Done**

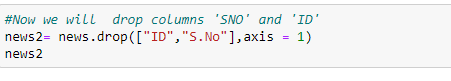
Now will check how many null value are their in the database and we will fill these null values.



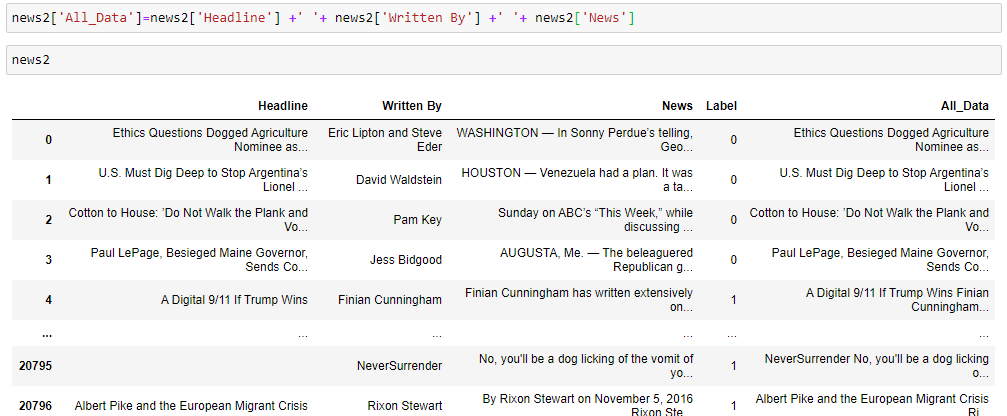
Now we will rename some of the columns name



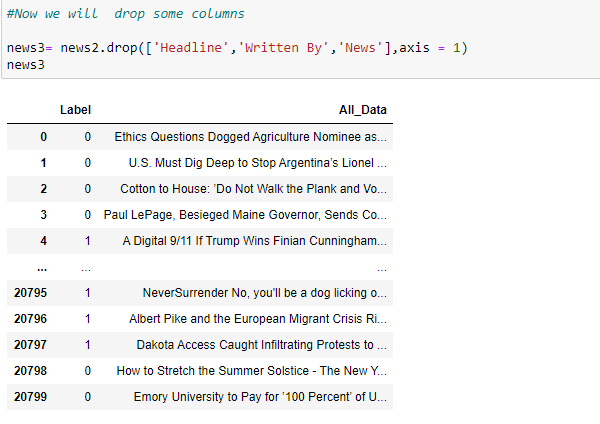
Now we will drop some of the columns which are not required in further data processing



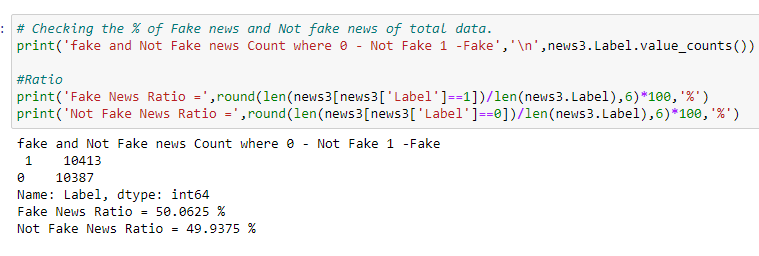
Now we will make a new column name ‘All\_Data’ which will contain all data of columns ‘Headline’ , ‘Written By’, ‘Label’ .



Now will drop some of the columns which are not required later on.



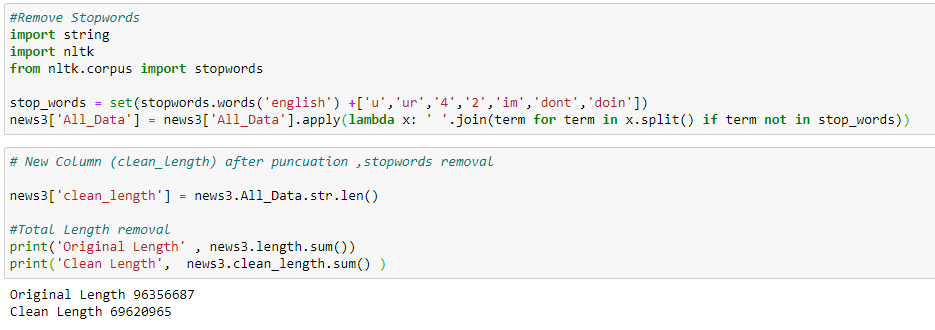
Now we will check % of ‘fake’ and ‘Not Fake’ news of the whole data.



Now will replace some of the words in database with some other words



Now we will remove stopwords and make database much more smaller and this will reduce time required for processing.



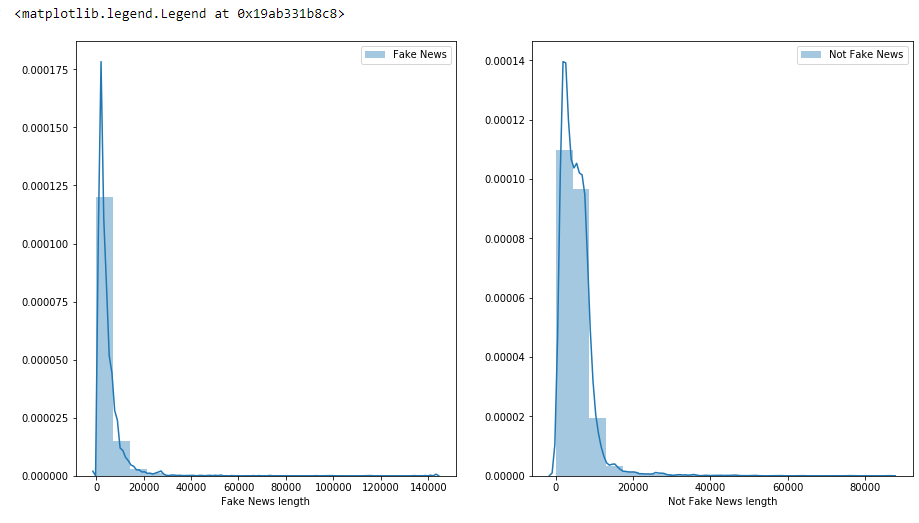
* Visualizations

Now we will make countplot of fake and Not fake news and will see their count.

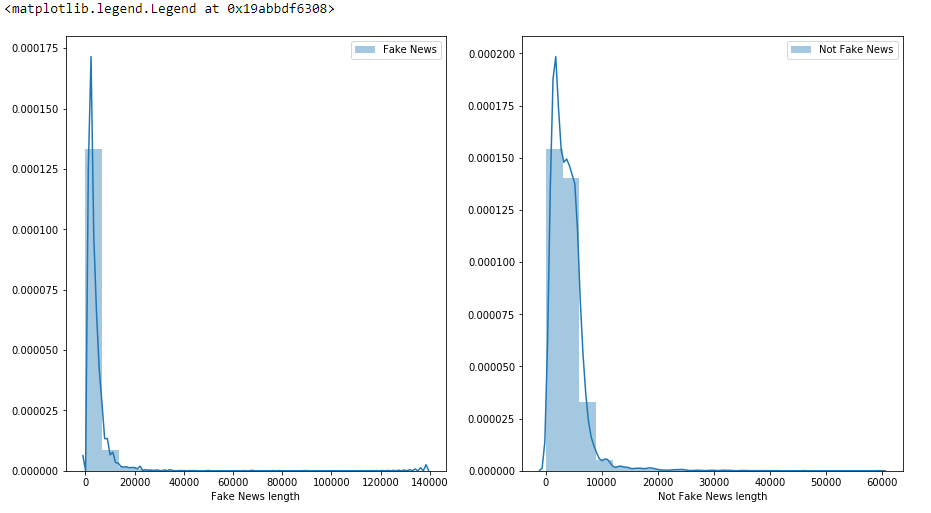


Now we will see distribution of the data before and after cleaning

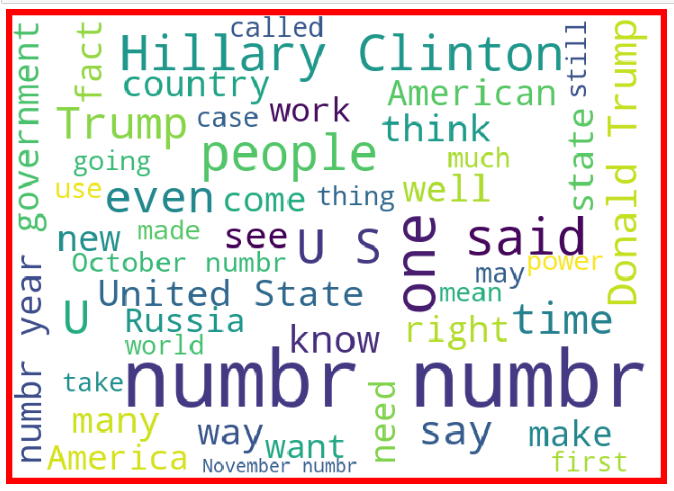
**Before cleaning**



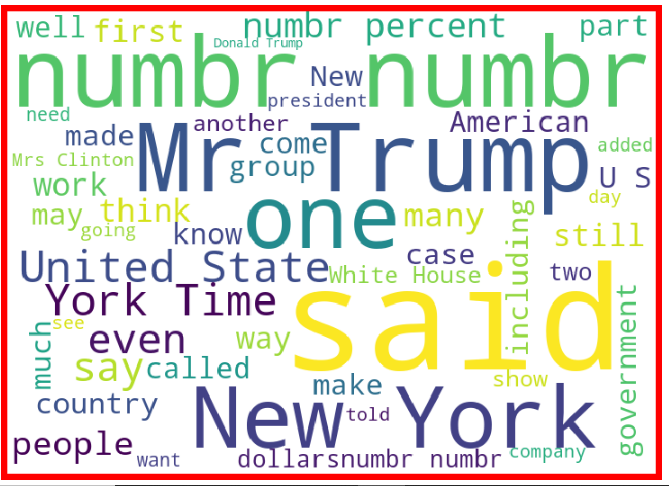
**After cleaning**



**Now we will see loud words in fake news**

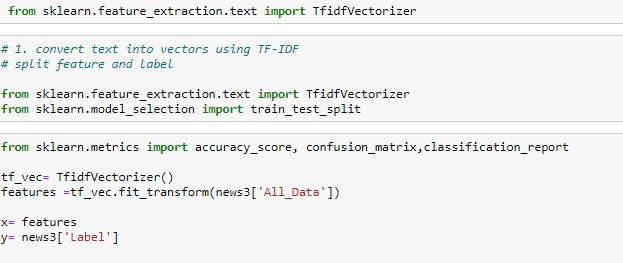


**Now we will see loud words in Not fake news**



* **Data Inputs- Logic- Output Relationships**

Now will divide the data into input and output , input will be ‘All\_Data’ column and output data will be ‘Label’ column and we will also import TfidfVectorizer to convert alphabets into vectors which will be required to build the model.



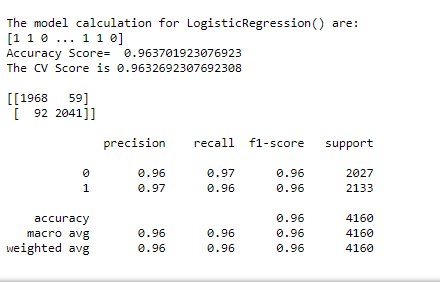
**Model/s Development and Evaluation**

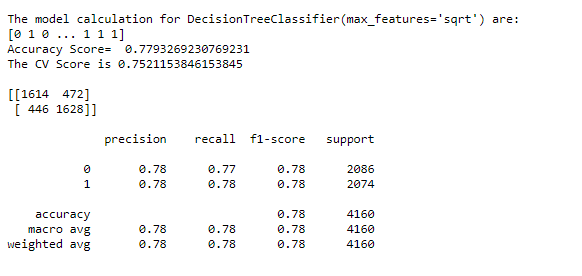
* **Identification of possible problem-solving approaches (methods)**

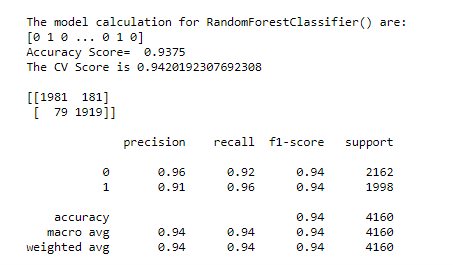
Since in this we have to predict whether news is fake news or not this is classification problem and we will use different algorithm like LogisticRegression , DecisionTreeClassifier , MultinomialNB, RandomForestClassifier ,AdaBoostClassifier etc.

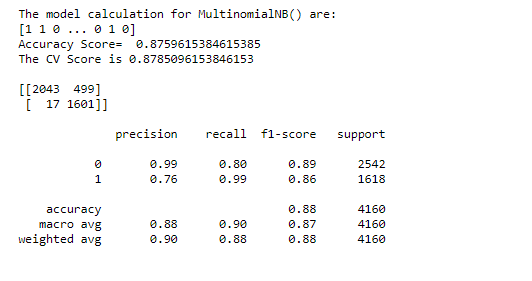
* **Run and Evaluate selected models**

Now we will run different model to check their performance and will check which model is performing best









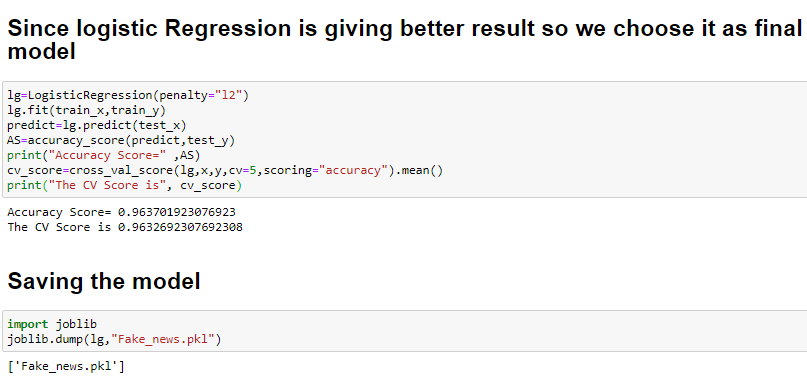
* Key Metrics for success in solving problem under consideration

In this we will select the model which has best accuracy score and precision , recall and f1 score .

**CONCLUSION**

* **Key Findings and Conclusions of the Study**

Of different algorithms we have used we found out that logistic regression is giving best accuracy score , cross value score , f1 score and precision score and lesser value False positive and false negative so will choose it as final model.



From the whole evaluation we can see that the maximum number of words are related to Trump and clinton and we can interpret that it was due to election campaign which was held during US presidential election and these fake news adversely effect the opinion of the voters . Most of these fake news related to trump were cleared by trump campaign but their was hardly any clarity or real news from the side of clinton and due to which it impacted the election result . Thus we can get idea how serious is Fake news Problem .

* **Learning Outcomes of the Study in respect of Data Science.**

With the help of the data science we have able to create the model

For the detection of the fake news using its different machine

Learning algorithm and NLP it give us a way to deal with the

complicated Problem and make our task simpler.