

**Fake News Detection**

**Submitted by:**

**Sourav Chaudhury**

**ACKNOWLEDGMENT**

I would like to take this opportunity to show my gratitude to the resources mentioned below for helping me complete this project successfully.

* Flip Robo Technologies
* DataTrained Education
* TowardsDataScience
* AnalyticsVidya
* StackOverflow
* GitHub

**INTRODUCTION**

* **Business Problem Framing**
* 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.
* Fake news is a major concern in our society right now. It has gone hand-in-hand with the rise of the data-driven era – not a coincidence when you consider the sheer volume of data we are generating every second!
* Fake news is such a widespread issue that even the world’s leading dictionaries are trying to combat it in their own way.
* **Conceptual Background of the Domain Problem**
* Fake news is an increasingly common feature of today’s political landscape. To help address this issue, researchers and media experts have proposed fake news detectors adopting natural language processing (NLP) to analyze word patterns and statistical correlations of news articles. While these detectors achieve impressive accuracy on existing examples of manipulated news, the analysis is typically quite shallow— roughly, models check whether news articles conform to standard norms and styles used by professional journalists. This leads to two drawbacks.
* First, these models can detect fake news only when they are under-written, for instance when the content is totally unrelated to the headline (so-called “clickbait”) or when the article includes words considered to be biased or inflammatory. While this criteria suffices to detect many existing examples of fake news, more sophisticated rumor disseminators can craft more subtle attacks, for instance taking a well-written real news article and tampering the article in a targeted way. By preserving the original subject matter and relating the content tightly to the headline without using biased phrases, an adversarial article can easily evade detection. To demonstrate this kind of attack, we evaluate a state-of-the-art model called Fakebox.
* **Review of Literature**
* Real news which is under-written or talks about certain political and religious topics is likely to be mistakenly rejected, regardless of its accuracy. This is a particularly serious problem for open platforms, such as Twitter in the United States and TouTiao in China, where much of the news is contributed by users with diverse backgrounds. To prevent frustrating false positives, platforms are still heavily relying on manual work for separating fake news from real news. We provide experimental evidence for Fakebox’s potential of misclassifying real news. Taken together, our experiments highlight vulnerable aspects of fake news detection methods based purely on NLP. Without deeper semantic knowledge, such detectors are easily fooled by fact-tampering attacks and can suffer from a high rate of false positives, mistakenly classifying under-written yet real news which may not be written in a journalistic style. To address these problems, we argue that some form of fact-based knowledge must be adopted alongside NLP-based models. The epidemic spread of fake news is a side effect of the expansion of social networks to

circulate news, in contrast to traditional mass media such as

newspapers, magazines, radio, and television. Human inefficiency to

distinguish between true and false facts exposes fake news as a

threat to logical truth, democracy, journalism, and credibility in government institutions. Veracity of information is an essential part of its integrity. The combat against fake news makes indissoluble the integrity and veracity checking of social networks’ information and data consumption in the application layer. The disclosure of fake content implies waste of processing and network resources. Further, it consists of a serious threat to information integrity and credibility of the provided service . Hence, the sharing of untrue information concerns the quality of Trust (QoT) applied to the news dissemination ,referring to how much a user trusts the content of a particular source. In different countries, it is possible to observe low levels of trust in the mass media, e.g., only 40% in the United States

whereas never-read links are highly shared (blindshares), e.g., 59% in the United Kingdom. In 2016, during the United States’ presidential elections, American society witnessed an alarming fake news epidemic, which had a multilateral effect. A similar effect also happened in the Brazilian elections in 2018. Due to its potential of dissemination, acceptance, and destruction , fake news is currently

one of the greatest threats to the concept of logical truth, having a high potential for deteriorating democracy, journalism, justice, and even economy. The economy, in particular, had to deal with fluctuations of 130 billion on the stock exchange as a result of a false statement claiming that Barack Obama had been injured in an explosion . In this context, there is a growing joint effort by the academic community to develop approaches that are capable of analyzing, detecting and intervening in the actuation of these misleading contents. Scientific evidence has already revealed the

vulnerability of humans to distinguish true from false. On average, human are correct 54% and, thus, our ability to identify fake and legitimate news is almost random. Aim to present the main algorithms and techniques that assist in linguistic characterization and detection of false news on social networks to guarantee the fake news term originally refers to false and often sensationalist information disseminated under the guise of relevant news. However, this term’s use has evolved and is now considered synonymous with the spread of false information on social media.

It is noteworthy that, according to Google Trends, the “fake news” term reached significant popularity in Brazil between the years 2017 and 2018, having its peak of popularity in October 2018, when there was the presidential election in Brazil.

Fake news is defined as news that is intentionally and demonstrably false, or as any information presented as news that is factually incorrect and designed to mislead the news consumer into believing it to be true. Sharma et al. argue that these definitions. Despite the lack of a clear consensus on the concept of fake news, the most accepted formal definition interprets news as intentionally and verifiably false. Regarding this definition, two aspects stand out: intention and authenticity. The first aspect concerns the dishonest intention of deceiving the reader. The second, on the other hand, relates to the possibility of this false information being verified. Fake news can be distinguished by the means employed to distort information. The news content can be completely fake, entirely manufactured to deceive the consumer, or it can be tricky content that employs misleading information to address a particular topic. There is also the possibility of imposing content that simulates genuine sources but, in fact, the sources are false. Other fraudulent characteristics of fake news content are the use of manipulated content, such as headlines and images that are not in accordance with the

content conveyed, or the contextualization.

* **Motivation for the Problem Undertaken**
* The growth of communications mediated by social media is one of the main factors that encourage the change of characteristics in current fake news. An individual’s inability to accurately discern fake news from the legitimate news leads to continued sharing and belief in false information on social media. It is difficult for an individual to differentiate between what is true and what is false while being overwhelmed with misleading information that is received over and over again. Furthermore, individuals tend to trust fake news because there is currently public disbelief in relation to traditional communication media. Additionally, the fake news is often shared by friends or confirms prior knowledge, which, for the individual, is more reliable than the discredited mass media. In this context, the identification of fake news is more critical compared to other types of information, since it is usually presented with elements that imbue it with authenticity and objectivity, thus making it relatively easier to obtain the public’s trust. Social media and collaborative information sharing on online platforms also encourage the spread of fake news, an effect called the echo chamber effect. The naive realism, in which individuals tend to believe more easily in information that is aligned with their points of view, the confirmation bias, in which individuals seek and prefer to receive information that confirms their existing points of view, and the theory of normative influence, in which individuals choose to share and consume socially safe options as a preference for acceptance and affirmation in a social group, are important factors in the perception and sharing of fake news that foster the effect of the echo chamber. These concepts imply the need for individuals to seek, consume and share information that is in line with their views and ideologies. As a consequence, individuals tend to form connections with ideologically similar individuals. In a complementary way, social network recommendation algorithms tend to personalize content recommendations that meet the preferences of an individual or group. These behaviours lead to the formation of echo chambers and filter bubbles, in which individuals are less exposed to conflicting points of view and are isolated in their own information bubble. The confinement of fake news in echo chambers, or information bubbles, tends to increase the survival and dissemination of such news. This is because the confinement incurs in the phenomenon of social credibility, which suggests that people’s perception of the credibility of information increases if others also perceive it as true, since there is a tendency for individuals to consider information to which they are submitted. Consume fake news on social networks. Due to the plurality of actors involved, the problem of identifying and mitigating the spread of fake news becomes even more complicated. The dissemination of fake news heavily relies on social media to the detriment of traditional media, due to the large scale, the reach of social media, and the ability to share content collaboratively. Social media websites have become the most popular form of fake news dissemination due to the increasing ease of access and popularization of computer mediated communication and Internet access. Concurrently, while in traditional journalism media, the responsibility of creating content remains with the journalist and the writing organization, moderation on social networks varies widely. Each social media is subjected to different moderation rules and content regulation. Information is consumed mainly by the general public or society, which constitutes an increasing number of social media users. The growth in the consumption of information through social media increases the risk of fake news causing widespread damage repeatedly as true.

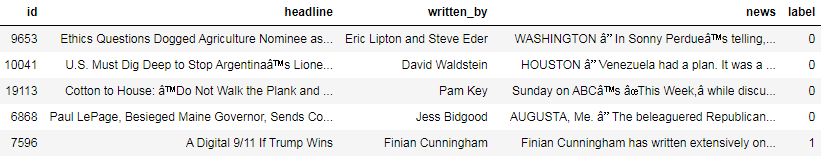
**Analytical Problem Framing**

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

In this project a machine learning model is constructed which is capable of classifying fake news from that of the real ones. An unclean dataset was given which was later cleaned using different strategies. Data preprocessing started with replacing missing values, then dropping columns followed by making required columns. Then a extensive feature engineering was done to reduce the data dimensionality. Visualizations are plotted to get a better understanding of the data and the problem statement. Text vectorization techniques are used to convert the text to numeric. Finally several algorithms are tested with the dataset in order to find the one that works best with the dataset.

* **Data Sources and their formats**

The data was provided in a CSV format. There were lot of missing values. The strategy that is considered here is taking values from the ones where the target variables were present. The column of the target variable is selected with the rows where there are values and accordingly they are selected. After filtering out the missing data of the label column below is the snapshot of the variables what were left with.

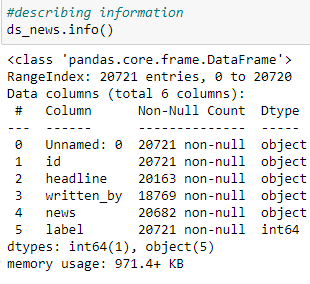


the above snapshot shows five columns in total where the column id will be dropped late as it is not important variable to consider while making the model.

* **Data Preprocessing Done**
* Data Preprocessing is one of the most important part of making any machine learning model. It involves processing the raw data in a way that it is easily understandable by the model for higher efficiency and efficacy. Below are some of the data preprocessing steps.

**Describing Information**

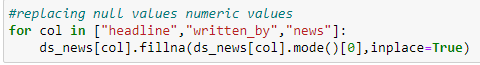
This step involves describing the variable names, the variable types and most importantly if there are missing values in the dataset.



The above table shows there are total of six columns with one numeric column and five categorical columns. There are three columns with missing values which needs to be replaced with data using required strategies.

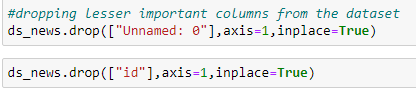
**Replacing Missing Values**

The above snapshot shows three columns with missing values. As all the columns with missing values are categorical in nature we will use the mode strategy to replace the missing values. Below is the code to do the same.



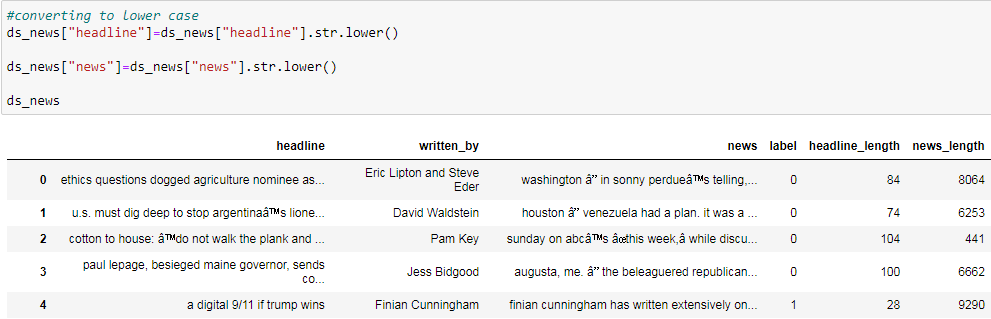
**Dropping Columns**

In this step the columns those are not important and will not be considered in the model will be dropped. Below are the codes to drop two columns from the dataset.



**Converting to Lower Case**

Converting the data to lower case will increase the models accuracy. It will be easier for the model to predict better in such scenarios. Below are the codes to convert the corpus to lower case.



**Feature Engineering**

A detailed feature engineering is performed to standardize the data and also to reduce the data dimensionality to make it more understandable and make better predictions by the model. The feature engineering performed with the dataset pertains to removing trailing spaces, white spaces and removing punctuations from the dataset. Below are the codes to perform the feature engineering.



The above feature engineering is performed in two columns i.e "news" and "headlines" that contains text elements.

**Removing Stopwords**

Stopwords are the most common words in any natural language. For the purpose of analyzing text data and building NLP models, these stopwords might not add much value to the meaning of the document.

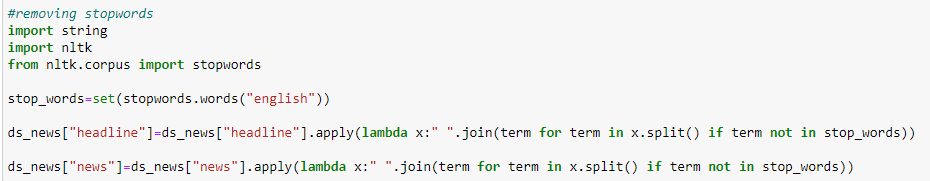
Here are a few key benefits of removing stopwords:

On removing stopwords, dataset size decreases and the time to train the model also decreases.

Removing stopwords can potentially help improve the performance as there are fewer and only meaningful tokens left. Thus, it could increase classification accuracy.

Even search engines like Google remove stopwords for fast and relevant retrieval of data from the database.

Below are the codes to remove the stopwords from the dataset.



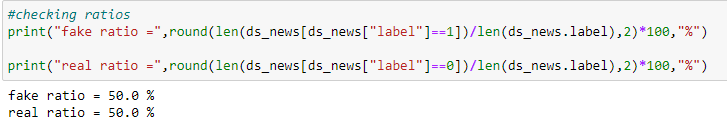
* **Hardware and Software Requirements and Tools Used**

This project is done using Python 3.0, Ms Excel and Ms Word. The GUI (Graphical User Interface) used in this project is Jupyter Notebook. The coding part of the project is done in Jupyter Notebook while the training and testing data is received in CSV format and the project report is written in Ms Word.

**Model/s Development and Evaluation**

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

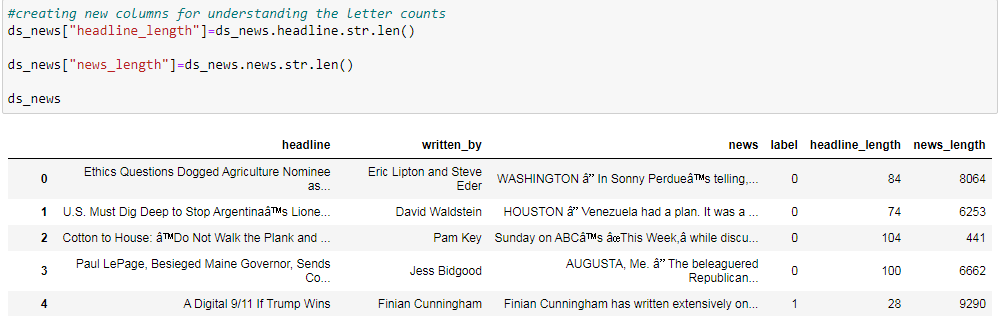
As this is a classification study, the target variables are binary in nature. We have labels in terms of 0s and 1s. The first thing that is laid down in understanding the problem statement understanding the distribution of the target variable. For this purpose, the ratios of the 0s and 1s are laid down. Below is the code for the same.



From the above ratios we can see that the 0s (Real News) and 1s (Fake News) have a similar ratios. This means our dataset is not imbalanced and we are good to go.

**Making New Columns**

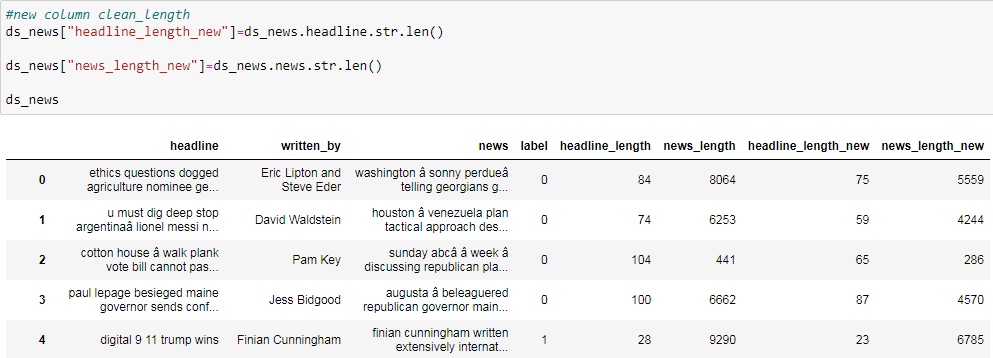
Two new columns are added in order to get a count of the letters in the text columns. This will help us to differentiate if the letter distribution of fake news are different than that of real news. Below is the code.

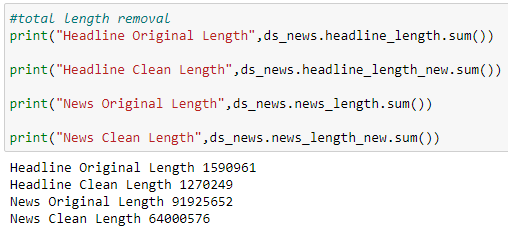


A above table shows there are two new columns displaying the count of the variables "headline and "news" characters.

**Length Removal**

The removal of stopwords has resulted in reduction of the data dimension. Two more columns has being added to find the new length of words and also to compare the length removal from that of the unclean data. Below is the code.



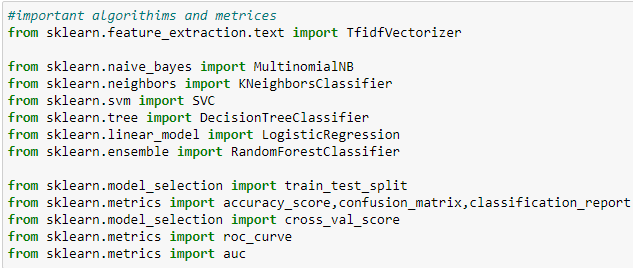


From the above code we can see that there is a huge difference from the cleaned data to that of uncleaned one.

* **Testing of Identified Approaches (Algorithms)**

**Importing Classification Algorithms and Metrics**

* Several algorithms are used to test the model and the one with the highest accuracy is finalized. Different classification metrics were also used to evaluate the models performance. Below are the list of algorithms and metrics used.



TfIdf vectorization technique is used to quantify the text to numeric data. Six classification algorithms are used to find the one which will be best suited for the dataset giving the higher accuracy. Train test split is used to separate the training and testing data.

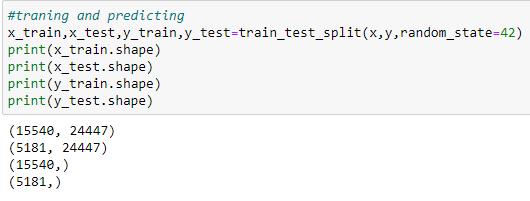
**Vectorization and Splitting Target and Input Variables**

TfIdf is used as a vectorize the text variables. There are two input variables "headline" and "news" and "label" is the target variable. A variable named features is made and the input variable is feeded into it and later it is treated as the "X" and the label is treated as "Y". Below is the code.



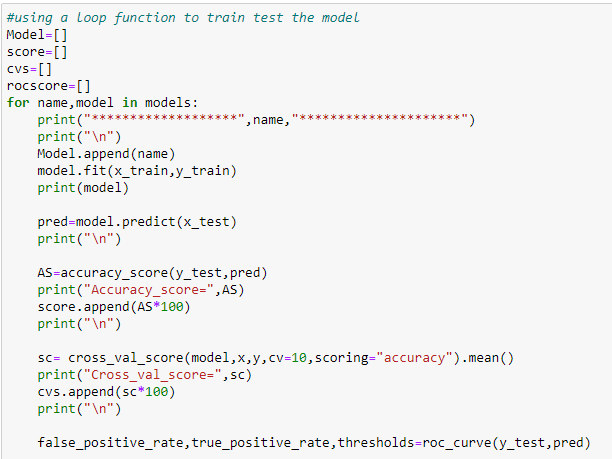
**Splitting Training and Testing Data**

Train test split is used to separate the training and testing data. Below is the code.



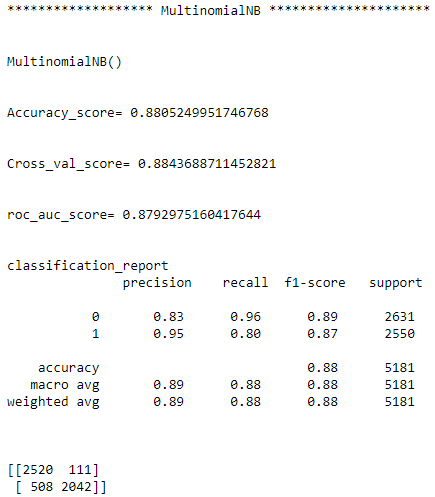
**Machine Learning**

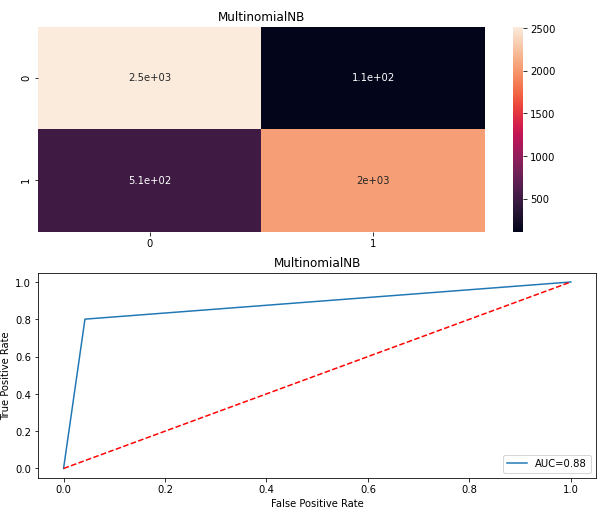
A loop function is used for the training and testing the dataset with six different algorithms. The one that performs the best will be finalized.

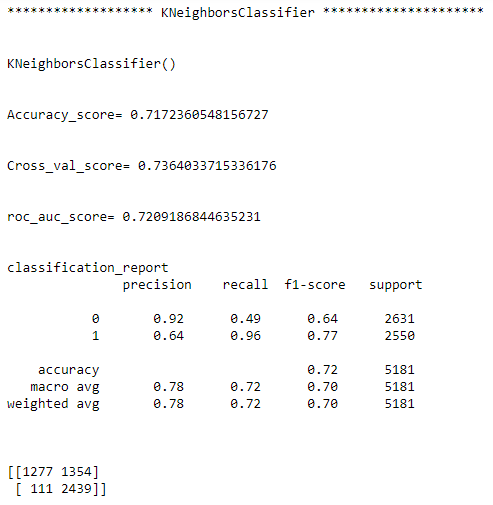
 

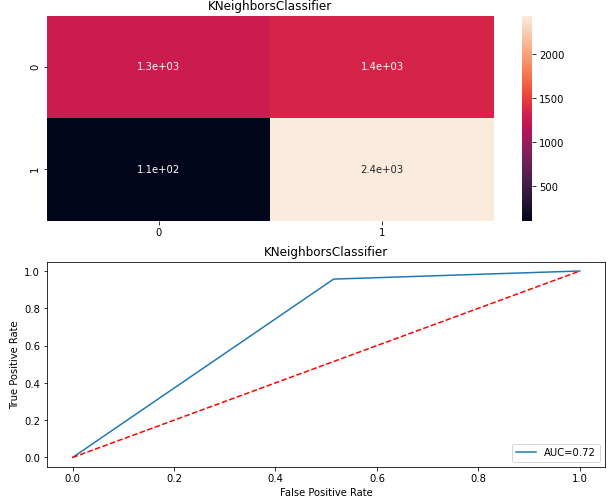
* **Run and Evaluate selected models**

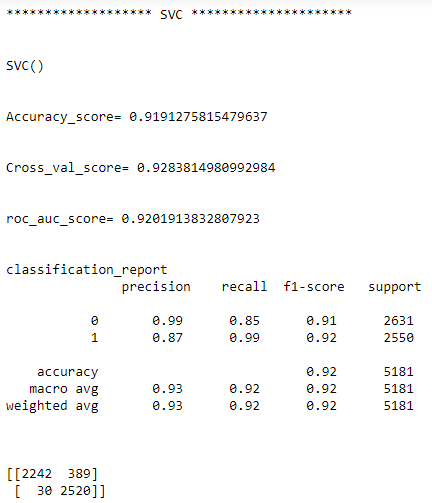
Below are the different algorithms with their scores based of various parameters.

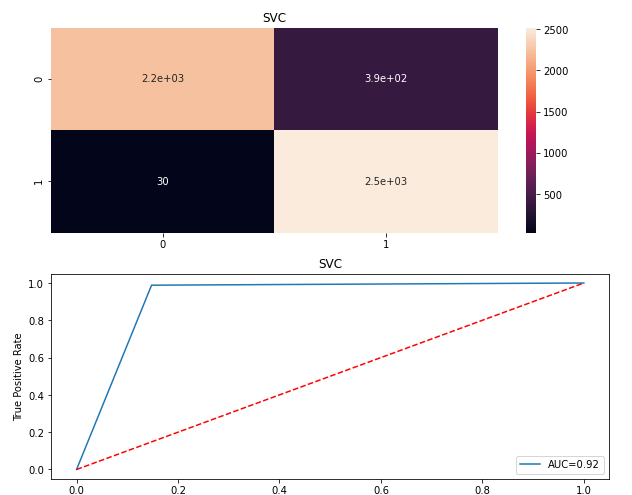


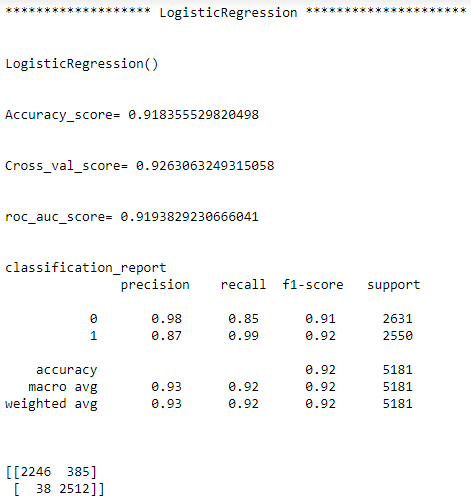


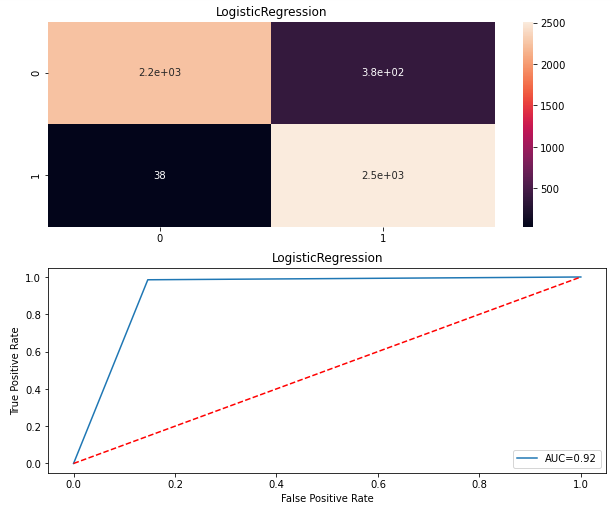


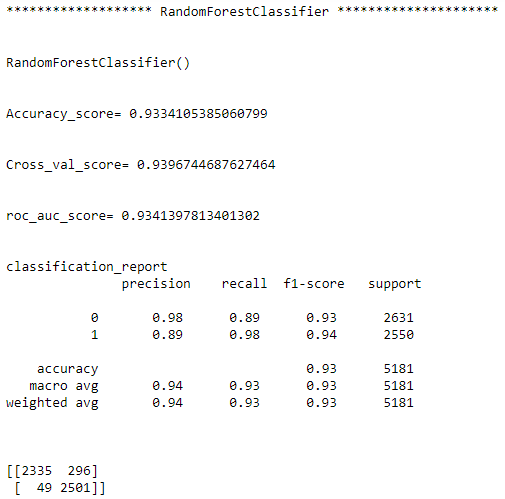


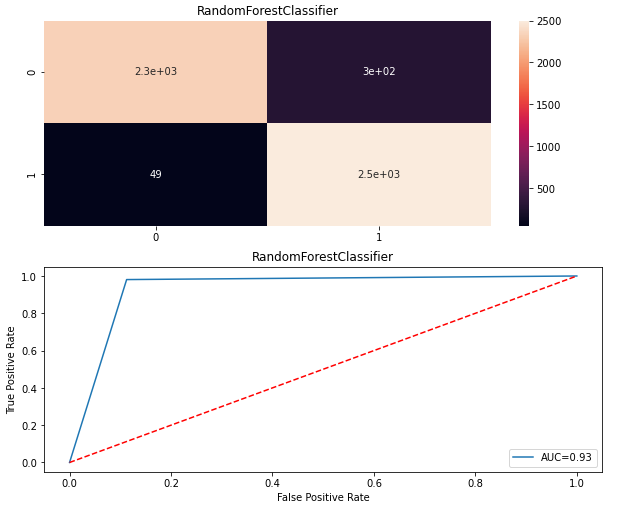




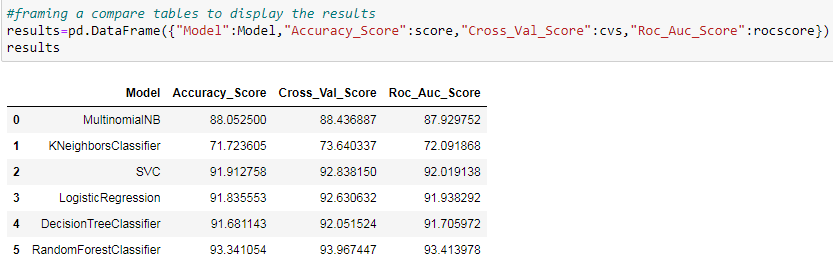




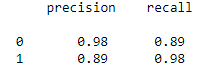
****

****

**Framing Compare Table**

****

The above shows a compare table which shows scores for the algorithms at once. We can see that random forest classifier has worked the best considering all the parameters. Also if we see the precision and recall for this algorithm, we can see that the this has worked the best considering the above parameters. Below is the snapshot.



* **Key Metrics for success in solving problem under consideration**

Metrics are various parameters that are used to judge the efficiency and efficacy of a model. The metrics used to judge different parameters of this model are below

* Accuracy Score
* Cross Validation Score
* ROC\_AUC Score
* Recall
* Precision
* F1 Score
* Confusion Matrix

**Accuracy Score**- Classification Accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples. It works well only if there are equal number of samples belonging to each class.

**Cross Val Score**- A key challenge with overfitting, and with machine learning in general, is that we can’t know how well our model will perform on new data until we actually test it. To address this, we can split our initial dataset into separate training and test subsets. There are different types of Cross Validation Techniques but the overall concept remains the same, to partition the data into a number of subsets, hold out a set at a time and train the model on remaining set and test model on hold out set

**ROC\_AUC Score**- The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

**Recall**- It is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).

**Precision**-  It is the number of correct positive results divided by the number of positive results predicted by the classifier.

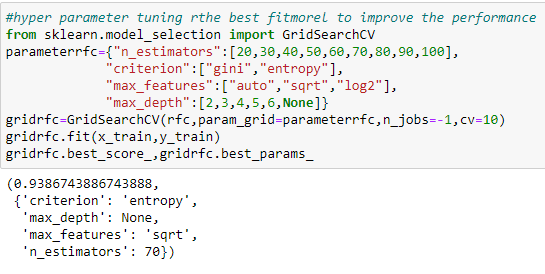
**F1 Score**- F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances). High precision but lower recall, gives you an extremely accurate, but it then misses a large number of instances that are difficult to classify. The greater the F1 Score, the better is the performance of our model.

**Confusion Matrix** - In the field of [machine learning](https://en.wikipedia.org/wiki/Machine_learning) and specifically the problem of [statistical classification](https://en.wikipedia.org/wiki/Statistical_classification), a confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) one (in [unsupervised learning](https://en.wikipedia.org/wiki/Unsupervised_learning) it is usually called a matching matrix). Each row of the [matrix](https://en.wikipedia.org/wiki/Matrix_(mathematics)) represents the instances in an actual class while each column represents the instances in a predicted class, or vice versa.

**Hyper Parameter Tuning**

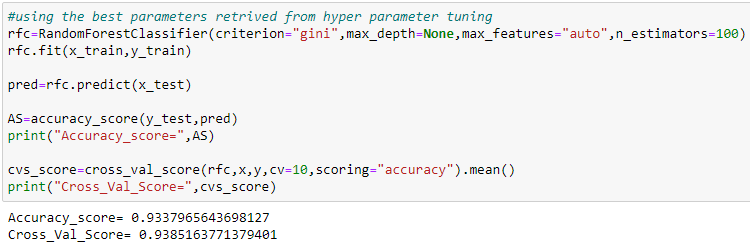
In  a hyper parameter tuning the [parameter](https://en.wikipedia.org/wiki/Parameter)s are tuned to control the learning process of the model. Different datasets with specific algorithms requires different learning rate for the model to work more accurately. So it forms different combinations to get the one parameter with best result. We will be tuning the Random Forest Classifier algorithm as it has the highest score.

Below is the code to perform hyper parameter tuning.



**Using Best Parameter**

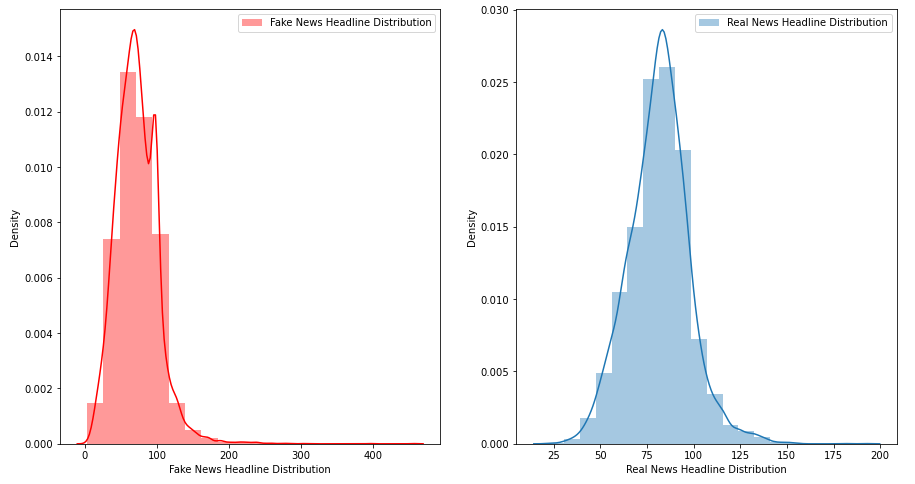
The above parameters are the best considering the dataset. We will be using the above parameters to generate the best score for the model. Below is the code.



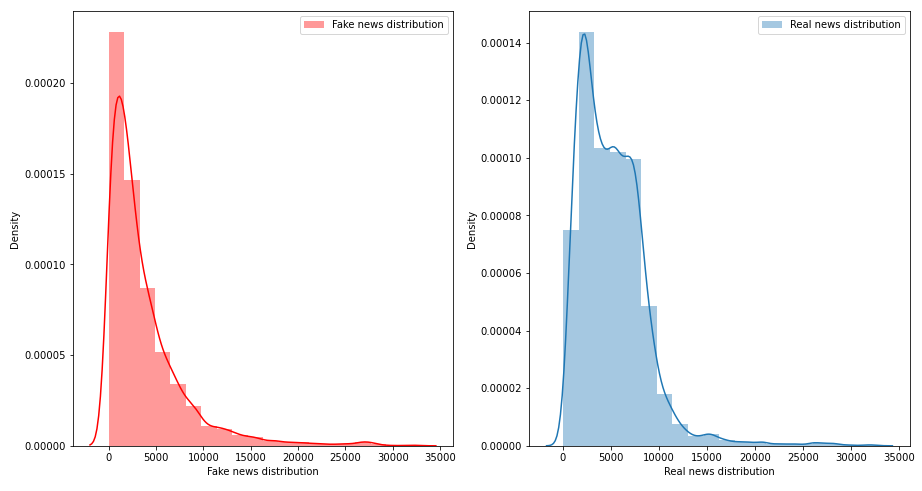
* **Visualizations**

**Plotting distribution of uncleaned length of words**

We have created two new columns for showing the length of letters for the fake news and the real news. This might be a good parameter for understanding if one can differentiate the fake news from that of the real news.

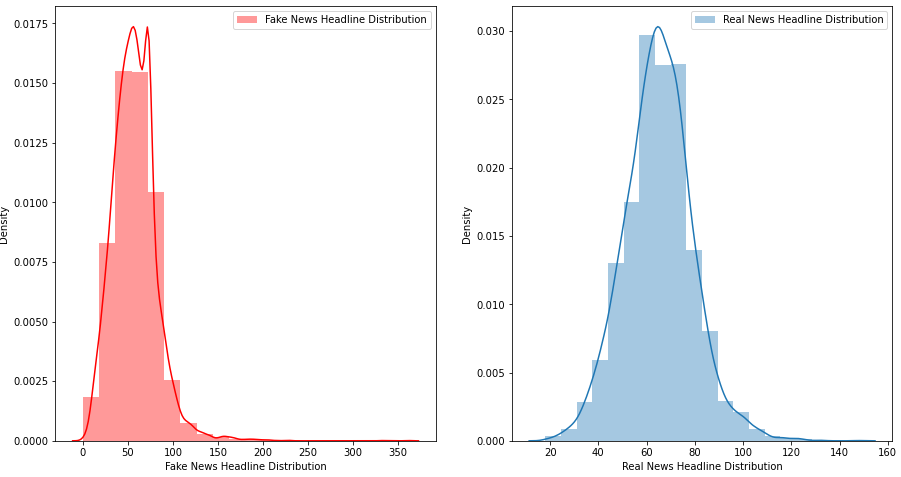


The above distplot shows the distribution of letters of the variable "headline" for both the fake news and the real news. We can see that for Fake news the letter distribution ranges from 0 to 400 where as for real news it ranges from 2 to 200.

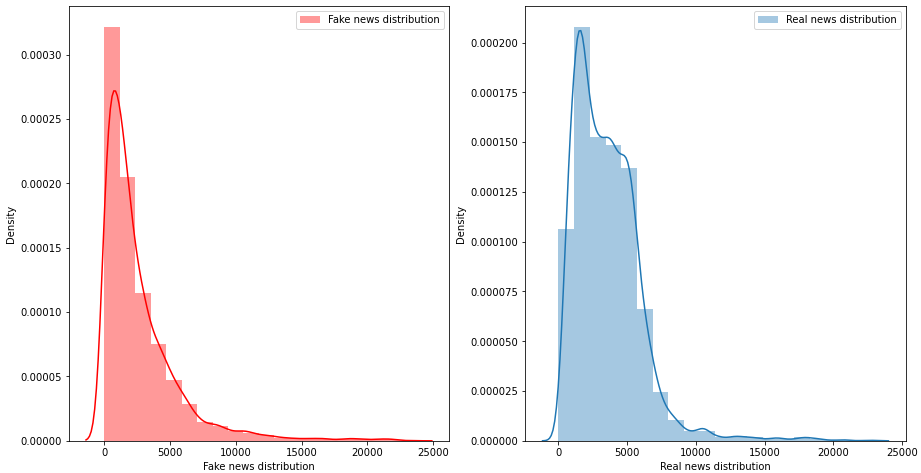


The above distplot shows the distribution of letters of the variable "news" for both the fake news and the real news. We can see that for Fake news the letter distribution ranges from 0 to 3500 and for real news also it is the same. We don't have much difference in this aspect.

**Plotting distribution of cleaned length of words**



The above distplot shows the distribution of letters of the variable "headline" for both the cleaned fake news and the cleaned real news. We can see that for Fake news headline the letter distribution ranges from 0 to 370 and for real news it ranges from 20 to 160.

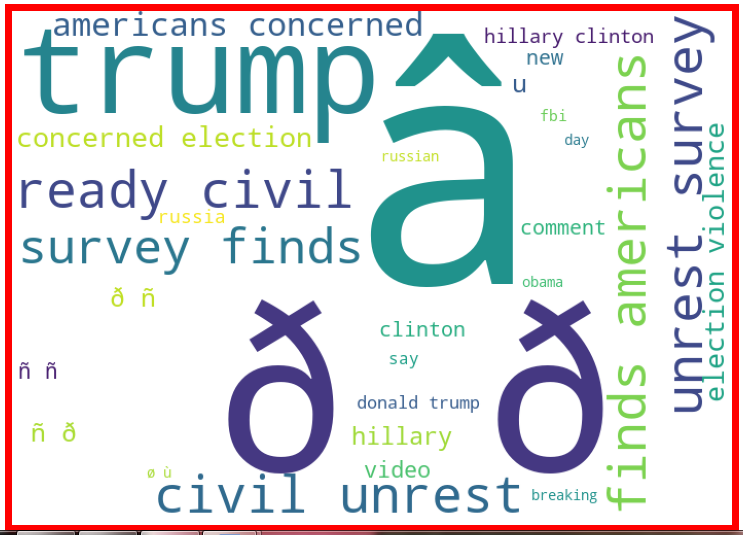


The above distplot shows the distribution of letters of the variable "news" for both the cleaned fake news and the cleaned real news. We can see that for Fake news article the letter distribution ranges from 0 to 2500 and for real news it ranges from 0 to 2500.

**Creating WordCloud**

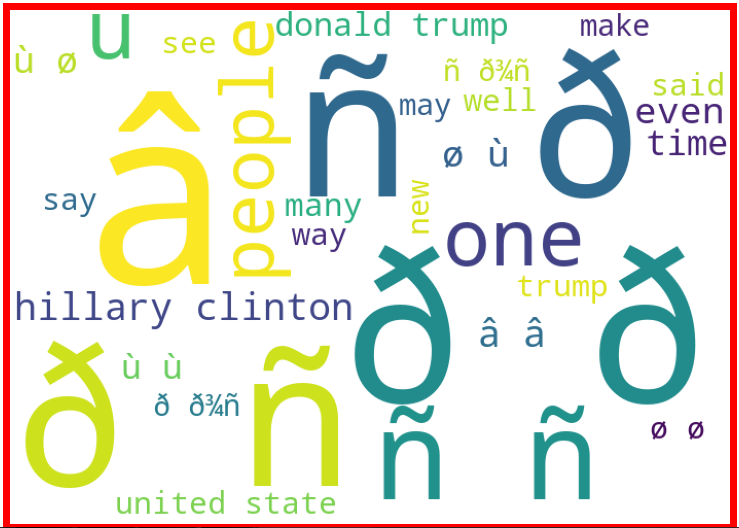
Word Clouds (also known as wordle, word collage or tag cloud) are visual representations of words that give greater prominence to words that appear more frequently. This type of visualization can help presenters to quickly collect data from their audience, highlight the most common answers and present the data in a way that everyone can understand.

**Creating word cloud for the fake news headline**

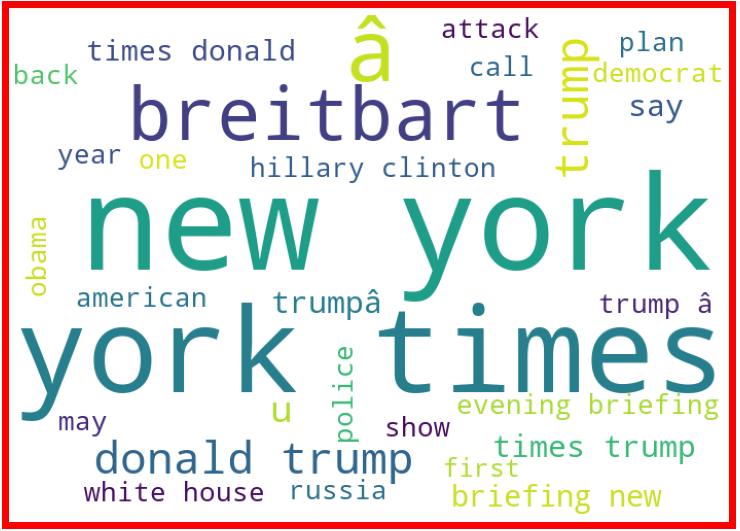


The above wordcloud is for the variable "headline" for fake news. From looking at the wordcloud we can find out that most of the repetitive words in the variable pertains to something that might be present in fake news itself.

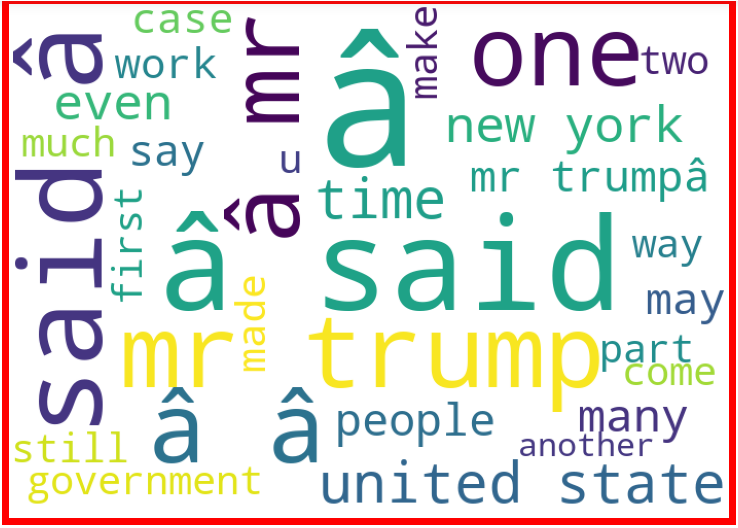
**Creating word cloud for the fake news body**

****

The above wordcloud is for the variable "news" for fake news. From looking at the wordcloud we can find out that most of the repetitive words in the variable pertains to something that might be present in fake news itself.



The above wordcloud is for the variable "headline" for real news. From looking at the wordcloud we can find out that most of the repetitive words in the variable pertains to something that might be present in real news itself.



The above wordcloud is for the variable "news" for real news. From looking at the wordcloud we can find out that most of the repetitive words in the variable pertains to something that might be present in real news itself.

* **Interpretation of the Results**

There are couple of visualizations that are made to have a clear and better understanding of the variables from the dataset. Firstly two new columns are added to display the length of two uncleaned columns "headline" and "news" and then two more columns are added to display the length of the cleaned columns "headline" and "news". A distplot is made to understand if there is any difference between the text length of fake news from that of the real news. The distplot shows the distribution of letters of the variable "headline" for both the fake news and the real news. We can see that for Fake news the letter distribution ranges from 0 to 400 where as for real news it ranges from 2 to 200. One more distplot shows the distribution of letters of the variable "news" for both the fake news and the real news. We can see that for Fake news the letter distribution ranges from 0 to 3500 and for real news also it is the same. We don't have much difference in this aspect. After this we made some word clouds to find the most occurring words for both real and fake news. The word cloud clearly showed differences between fake news from that of real ones. Six algorithms are used to test and train the model and Random Forest Classifier had the best score out of all. After this hyper parameter tuning is performed to generate the best parameters and then those are used to generate the best scores for the model.

**CONCLUSION**

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

An uncleaned dataset came which later was cleaned using techniques. The strategy that is used in this case is that all the blank rows from the target variable are omitted and only rows with valid information from the target variable are considered. This let us in capturing the dataset with valid information for all the columns. After this step there were still some rows with null values. As all the columns with null values are categorical in nature, we have used mode strategy to replace the missing values. The labels were binary in nature so we checked for class imbalances. The % of 0s and 1s in the dataset were equal. Then the entire dataset was converted to lower case to reduce the data dimensionality. This is a great technique to make the data uniform in nature and also make it more readable by the model, which will result in increasing the efficacy of the model. A detailed feature engineering is also performed to reduce the data dimensionality. It included removing trailing spaces, white spaces, single spaces and punctuation. Also a detailed removal of stopwords is performed. New columns are made to display the text elements of the variables to differentiate fake news from real ones. Visualizations such as wordcloud distribution plot are made to get a deeper understanding of the data. Finally six algorithms are used to train and test the data and the one with the highest score is later tuned to get the best parameters which again is used to generate the best score for the model.