

Fake News Classifier

Submitted by:

Prince aggarwal

**ACKNOWLEDGMENT**

I would like to express my gratitude to all those who provided me the opportunity to take a step in this project. I gone through google and Kaggle for the references and my old projects helps me understanding the way to solve this use case.

**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.

In this modern and digital era, the possibility of fake news is increased. So, there are various company which are investing big in classifying the fake news because one fake news can make a big loss to industry and lead to worse circumstance for the people.

* Conceptual Background of the Domain Problem

To check the source of the news and avoid making judgement on single news better is to collect same news from different sources and then come to any decision.

* Review of Literature

Deep learning will be very impactful method for this project .RNN - Using NLP with Machine Learning because while using NLP we came to know more insight about the data by doing text pre-processing and come to suitable conclusion.

* Motivation for the Problem Undertaken

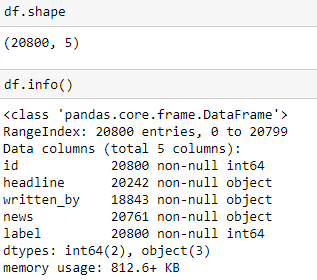
Because of the most demanding and most exploratory technology i.e. NLP and specifically using machine learning in this type of problem can help me to learn more and explore me more. The following problem statement is also a useful use case because as said earlier fake news classification is becoming more important now a days.

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem

1. Data set contain 20800 rows which is a great number followed by 4 main columns.
2. All the fields are in text form except id , Unnamed:0 and datatype is object for the main columns/features.
3. Dataset contain null values which are very less in numbers, so they can be directly drop the null values from the dataset.

* Data Sources and their formats



So above image described the dataset i.e. total features,shape of dataset total count, their types and memory usage of the dataset.

* Data Pre-processing Done

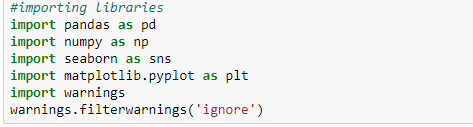
1. For Data Cleaning, there are 2 features in the dataset “unnamed:0” and “id” which are not playing an important role for the model building, so got removed them from the dataset.
2. There are some null values in the dataset in columns “headlines”, ”news” and “written\_by”
3. “written\_by column is not playing important role in this process and contains null values therefore it is removed
4. Null values are dropped as they were very small in range as compared to dataset
5. Also, while doing text pre-processing, I used TFIDF vectorizer

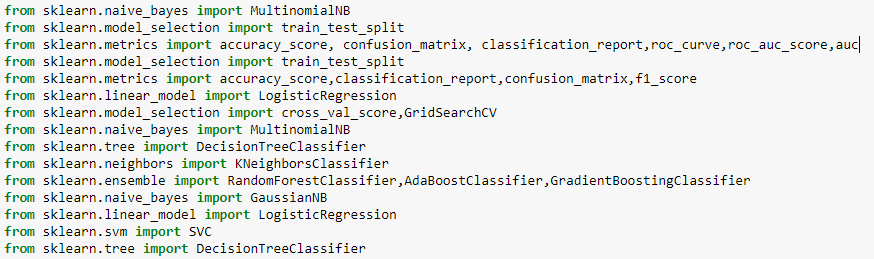
* Data Inputs- Logic- Output Relationships

The inputs are object data i.e. the data is in corpus form and needs to be in vector form for models.



* Hardware and Software Requirements and Tools Used





Above are the libraries which I used to pre-process, predict and visualize the project.

Pandas - pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

Seaborn, Matplotlib – This is a visualization library which helps to plot different type of bar charts and line charts which help to get more information from the dataset.

Sklearn – Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

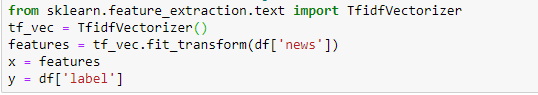
I followed the path of lifecycle of Data Science which includes:

Diagram

Description automatically generated

1. Read the use case and search from the google to know more about the use case and in which field or domain it will applicable more.
2. As data set is already there, so this step is excluded.
3. Data Preparation includes the data cleaning and describing the data which I followed and removing the null values.
4. EDA involves the visualization which is helpful to get more insight from the data and get to know about the trend of features individually and in group and check the most frequent words.
5. Modelling involves creating the model with suitable algorithm which provide the best result, I tried multiple algo and apply hyperparameter tuning.
6. Model Evaluation, for this I used confusion metrices and mainly focus on False positive and tried to reduce the False positive which is type 2 error and plot ROCAUC curve which is also covering most of the area under curve.

* Testing of Identified Approaches (Algorithms)



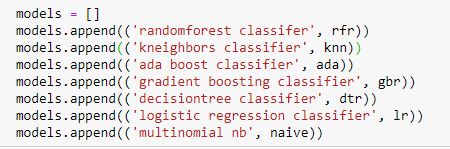


Above are the algorithms used in this model for training and testing

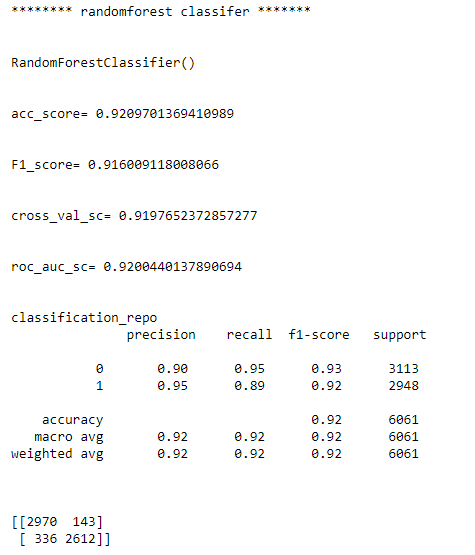
Tfidf is used to convers text to vector

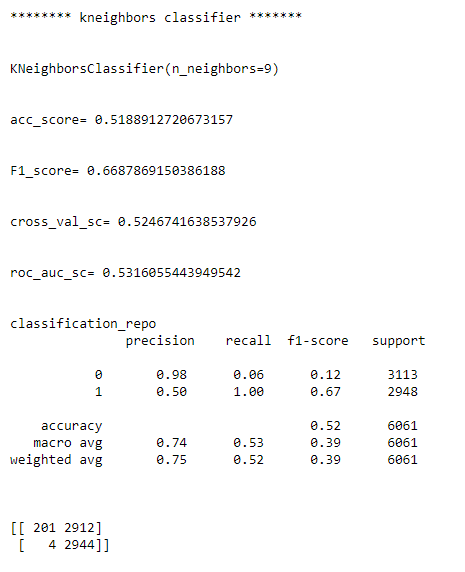
And rest algorithms are perfomed on various metrices

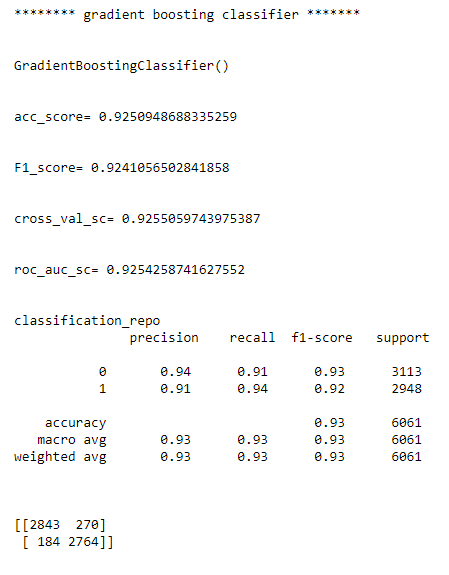
* Run and Evaluate selected models

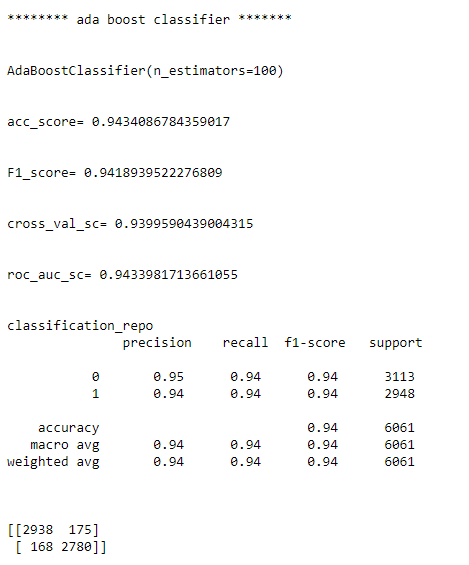


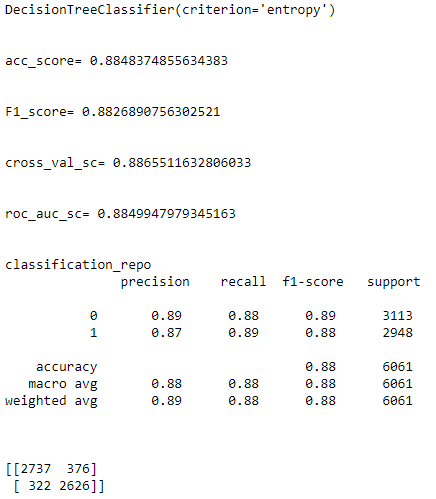
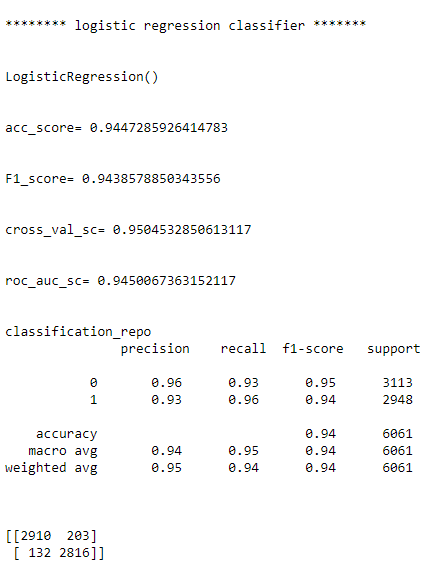


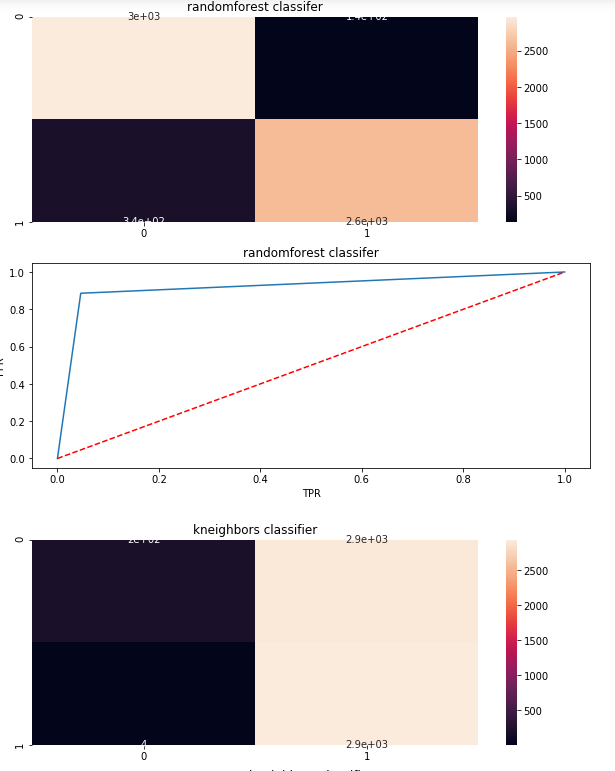
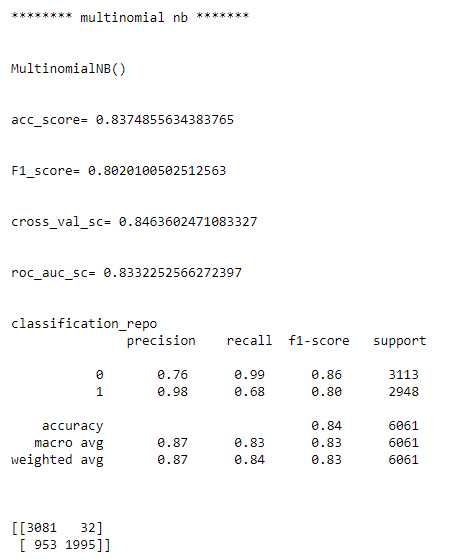


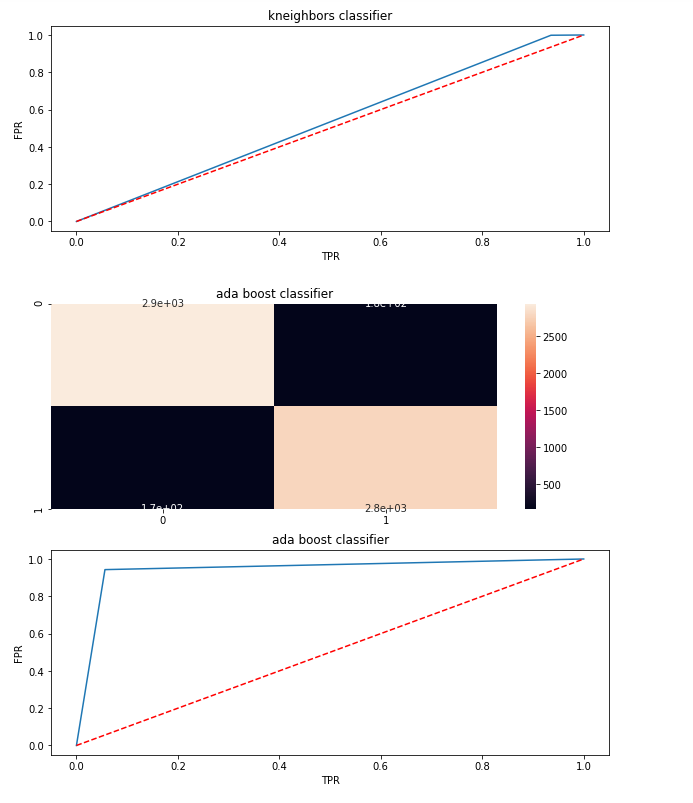


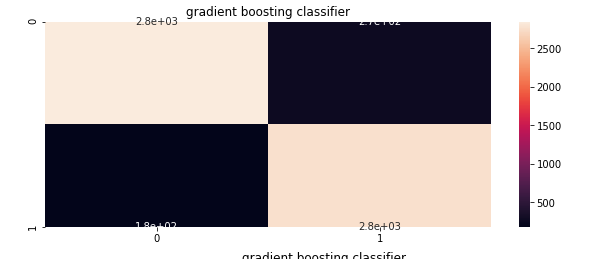


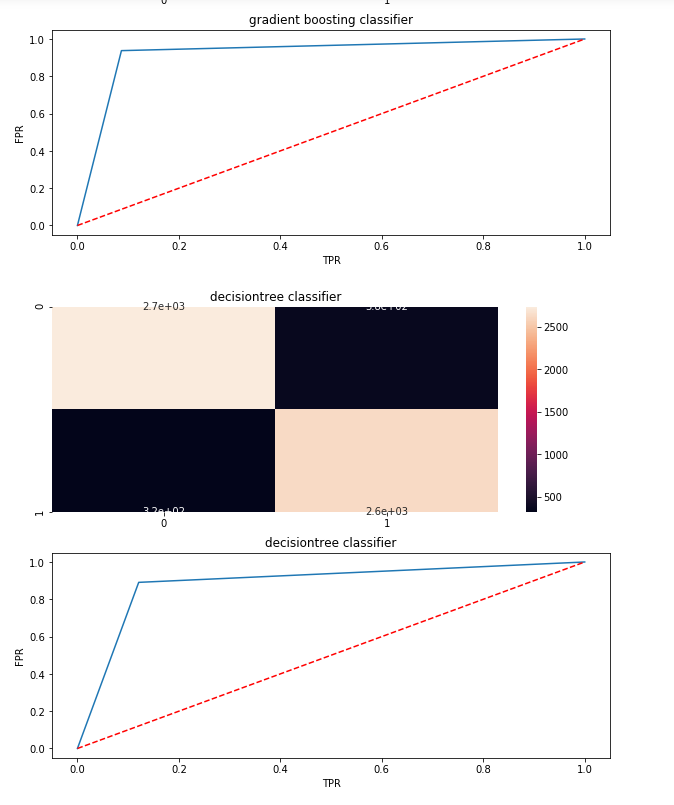


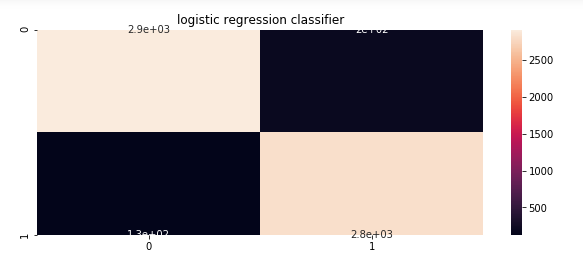


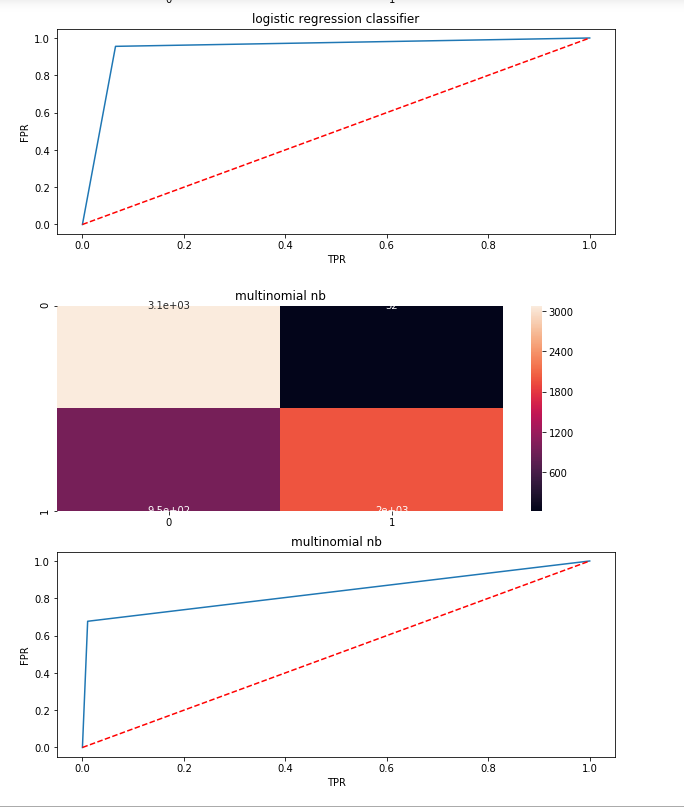




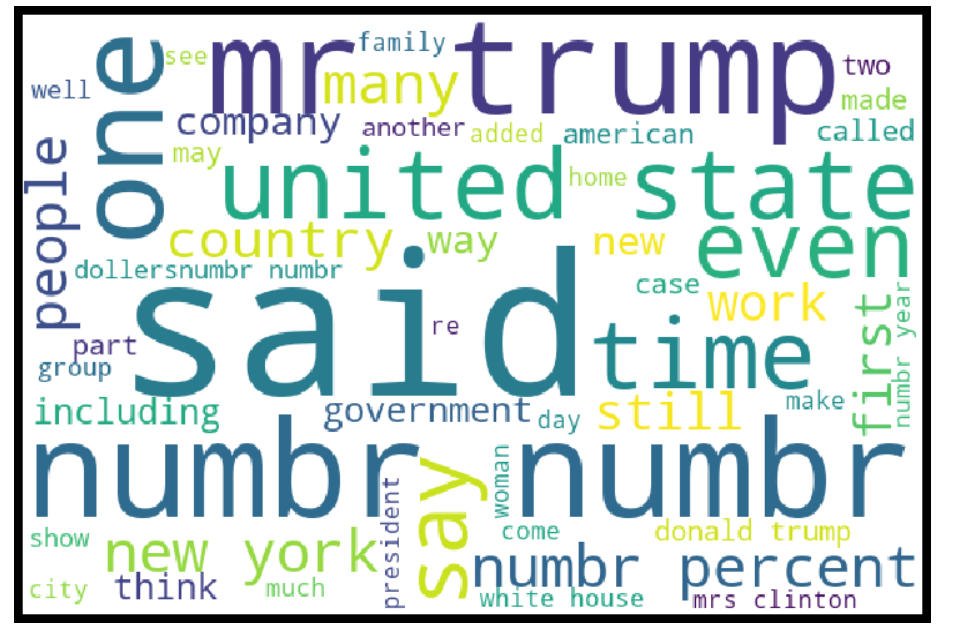




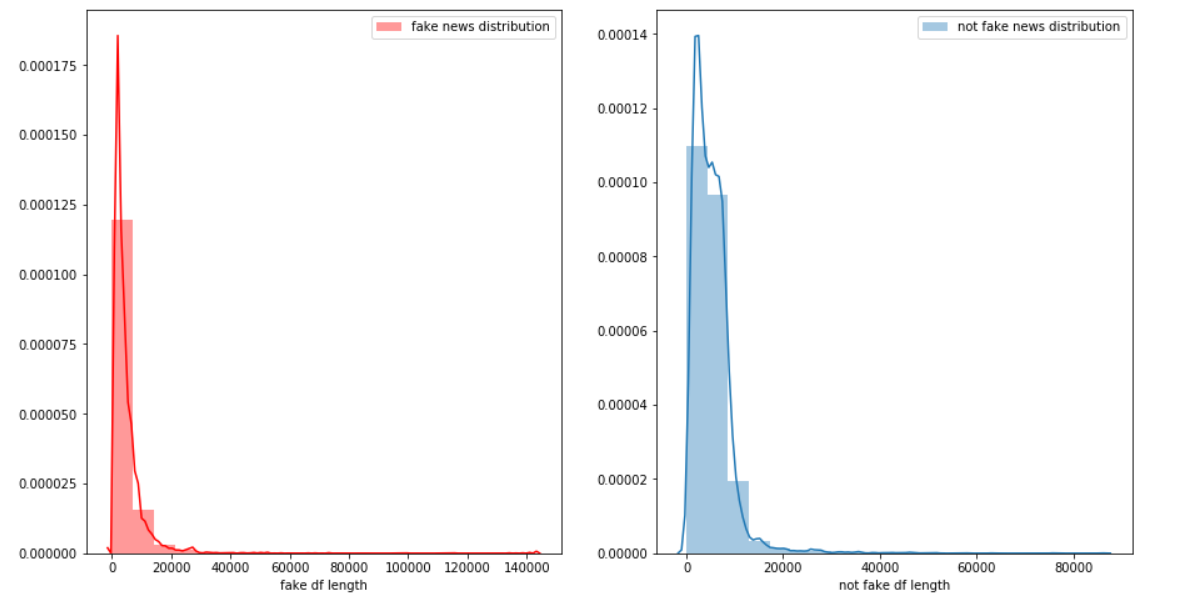




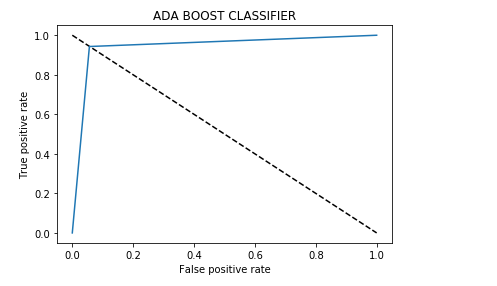
Visualizations

**Word Cloud:** ****

**Length distribution distplot :**



**Roc Curve:**



Above graph clearly show the sign of good model as most of the area is lying under the graph and score is also near to 90%

**CONCLUSION**

* Key Findings and Conclusions of the Study
* There are more than 60% news that are real.
* Most of the real news are containing words followed by election, president etc.
* Removing the column that are unnamed ,id , written\_by , headline does not impact the model training.
* Also, using Tfidf vectorizer along with hyper parameter tuning reduce the false positive error.
* Using the roc-auc-curve most of the region fall under the curve and accuracy is near to 94%.
* From multiple model it is shown that Ada boost with Tfidf suit well for the data and have better performance as compared to others models and there is no under/overfitting .
* Learning Outcomes of the Study in respect of Data Science
* While implementing this project most of the time is taken doing the text pre-processing and finding the best parameters for the models .
* Gain knowledge and get more insight of various stemmer and vectorizer.
* In my finding, ada boost classifer , # ensemble classifiers are well suitable for large dataset specially in text data, because of concept of Ada Boost is an iterative ensemble method. Ada Boost classifier builds a strong classifier by combining multiple poorly performing classifiers so that you will get high accuracy strong classifier. The basic concept behind Ada boost is to set the weights of classifiers and training the data sample in each iteration such that it ensures the accurate predictions of unusual observations
* Limitations of this work and Scope for Future Work
* There should be more rows and data so that it will work well, and performance of model would increase.
* Also, it has future scope in various use cases likewise in election, social media etc, where every day there are multi fake news spread.
* So, in the future it may use very well to easily classify the news as fake or real.