

FAKE NEWS CLASSIFICATION

Submitted by-

Madhurima Srivastava

Internship 30

Acknowledgment

Foremost, I would like to express my sincere gratitude to Data Trained team for the continuous support of my Data Science study and research, for the patience, motivation, enthusiasm, and immense knowledge. The guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Data science study.

Besides Data Trained, I would like to thank Flip Robo Team, for their encouragement, insightful internship, and help to understand the study.  
 My sincere thanks also go to SME Mohd Kashif for offering me the internship opportunities in their ally and leading me working on diverse exciting projects. I am over helmed in all humbleness and gratefulness to acknowledge my depth to all those who have helped me to put these ideas, well above the level of simplicity and into something concrete.

Last but not the least, I would like to thank my family, my parents, for being a morale support to me at the first place and backing me up in any down time, for being there in all hard situations making me to fight against any difficulty I face.

Introduction

## Business Problem Framing

Fake news is false or misleading information presented as news. It often has the aim of damaging the reputation of a person or entity, or making money through advertising revenue. However, the term does not have a fixed definition, and has been applied more broadly to include any type of false information, including unintentional and unconscious mechanisms, and also by high-profile individuals to apply to any news unfavourable to his/her personal perspectives.

Once common in print, the prevalence of fake news has increased with the rise of social media, especially the Facebook News Feed. Political polarization, post-truth politics, confirmation bias, and social media algorithms have been implicated in the spread of fake news. It is sometimes generated and propagated by hostile foreign actors, particularly during elections. The use of anonymously-hosted fake news websites has made it difficult to prosecute sources of fake news for libel. In some definitions, fake news includes satirical articles misinterpreted as genuine, and articles that employ sensationalist or clickbait headlines that are not supported in the text.

Fake news can reduce the impact of real news by competing with it; a Buzzfeed analysis found that the top fake news stories about the 2016 U.S. presidential election received more engagement on Facebook than top stories from major media outlets. It also has the potential to undermine trust in serious media coverage. The term has at times been used to cast doubt upon legitimate news, and former U.S. president Donald Trump has been credited with popularizing the term by using it to describe any negative press coverage of himself. It has been increasingly criticized, due in part to Trump's misuse, with the British government deciding to avoid the term, as it is "poorly- defined" and "conflates a variety of false information, from genuine error through to foreign interference".

Multiple strategies for fighting fake news are currently being actively researched, and need to be tailored to individual types of fake news. Effective self-regulation and legally-enforced regulation of social media and web search engines are needed. The information space needs to be flooded with accurate news to displace fake news. Individuals need to actively confront false narratives when spotted, as well as take care when sharing information via social media. However, reason, the scientific method and critical thinking skills alone are insufficient to counter the broad scope of bad ideas. Overlooked is the power of confirmation bias, motivated reasoning and other cognitive biases that can seriously distort the many facets of immune mental health. Inoculation theory shows promise in designing techniques to make individuals resistant to the lure of fake news, in the same way that a vaccine protects against infectious diseases.

## 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.

In the below blog we are going to see about how we are classifying the fake news with the genuine news; we are going to use several machine learning techniques and we will plot and analyse how to identify a news as fake. I have tried using several NLP techniques and arrived at a model that will classify news is fake or genuine.

## Review of Literature

### The purpose of the literature review is to:

1. Identify the News basis on the content and author to tell weather it is fake or not
2. Stop the spread of fake news which will potentially spread incorrect information amount the people.

To solve this problem, we are now building a model using our machine learning technique that identifies all the Fake news, using the same the news companies can avoid the spread of fake news in all the mediums.

I have used 8 different Classification algorithms and shortlisted the best on basis on the metrics of performance and I have chosen one algorithm and build a Machine Learning model in that algorithm.

## Motivation for the Problem Undertaken

Fake news is a topic that has gained a lot of attention in the past few years, and for good reasons. As social media becomes widely accessible, it becomes easier to influence millions of people by spreading misinformation. As humans, we often fail to recognize if the news we read is real or fake. A study from the University of Michigan found that human participants were able to detect fake news stories only 70 percent of the time. But can a neural network do any better? Keep reading to find out.

### The goal of this article is to answer the following questions:

* What kinds of topics or keywords appear frequently in real news versus fake news?
* How can we use a deep neural network to identify fake news stories?

**DATA SOURCES AND THEIR FORMATS**

* You can find many datasets for fake news detection on Kaggle or many other sites. This is datasets from Kaggle. There are two datasets one for fake news and one for true news. In true news, there is 21417 news, and in fake news, there is 23481 news. You have to insert one label column zero for fake news and one for true news. We are combined both datasets using pandas built-in function.

**Some Useful Links**-

1-https://www.javatpoint.com/nlp

2-https://www.educative.io/answers/preprocessing-steps-in-natural-language-processing-nlp

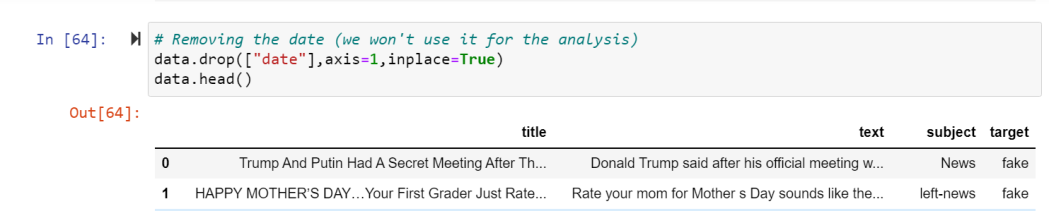
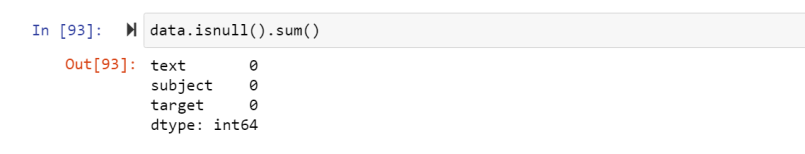
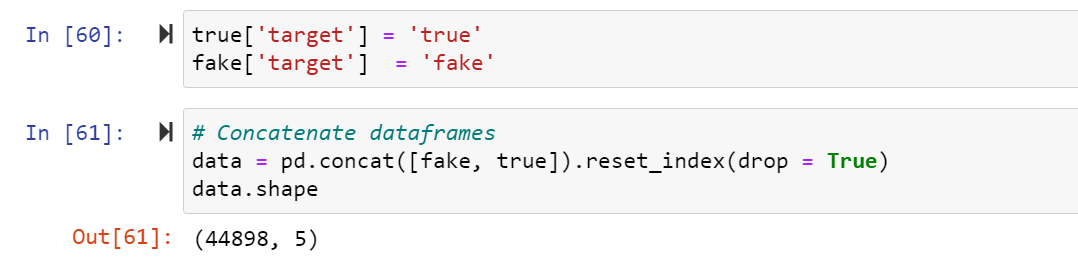
3https://www.youtube.com/watch?v=5ctbvkAMQO4

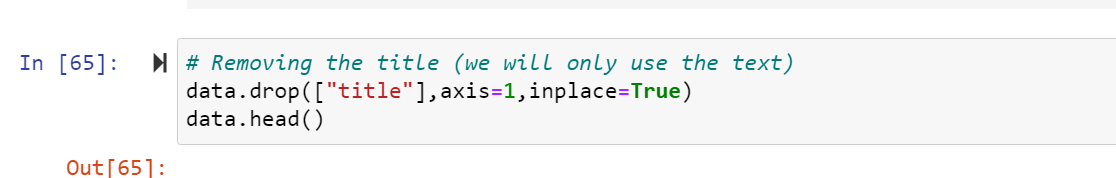
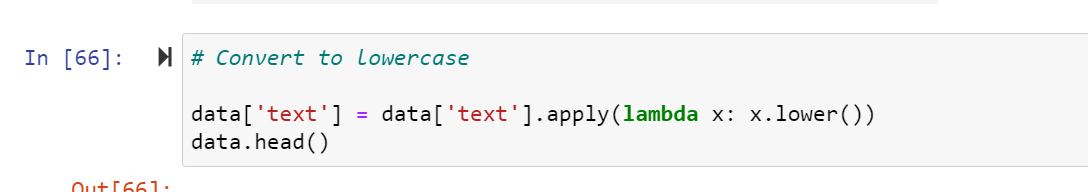
4https://www.youtube.com/watch?v=X2vAabgKiuM

**DATA** **PRE-PROCESSING**

**Checking** **the** **missing** **values**

Moving ahead, cleaning the dataset will remove errors which in turn will increase productivity and render highest quality information in decision making. Here, in this dataset, the null values are checked using isnull().sum(), which results in getting 2 features having null values in actual.

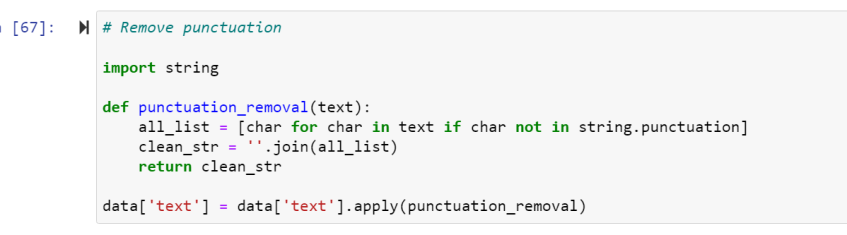
**Dropping** **unnecessary** **data**

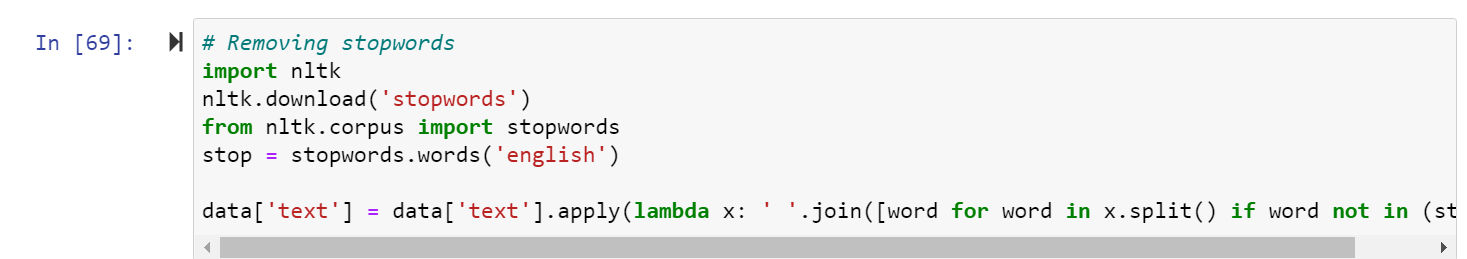


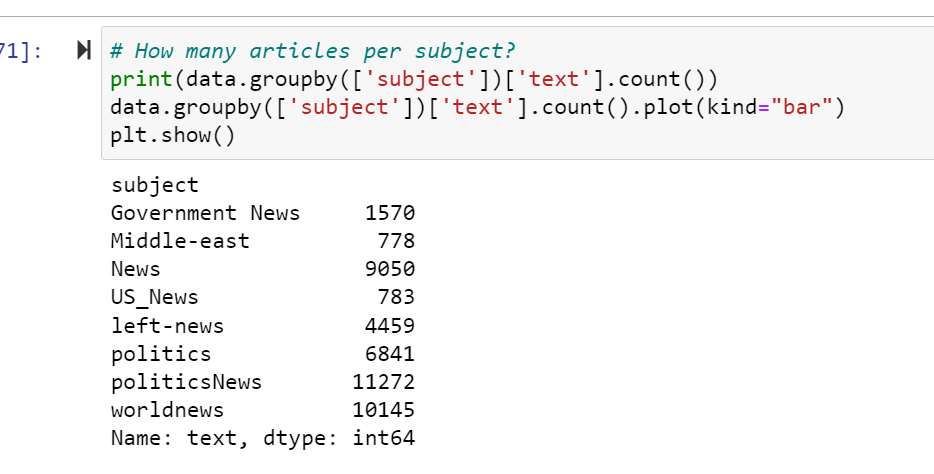
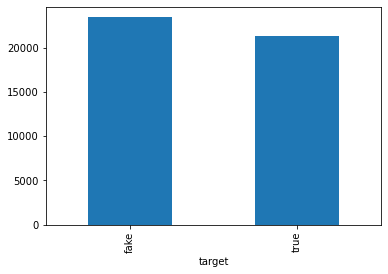
**Preprocessing** **using** **NLP**

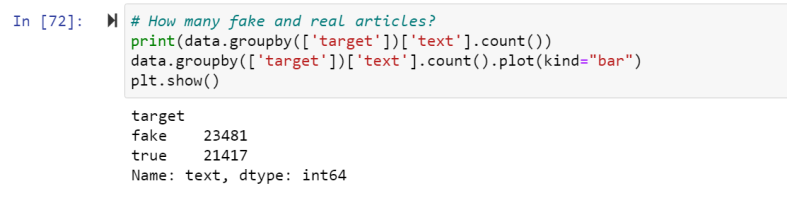
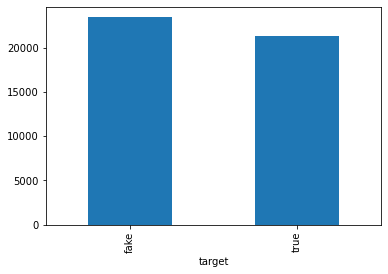
Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. This data is usually not necessary or helpful when it comes to analysing data because it may hinder the process or provide inaccurate results.

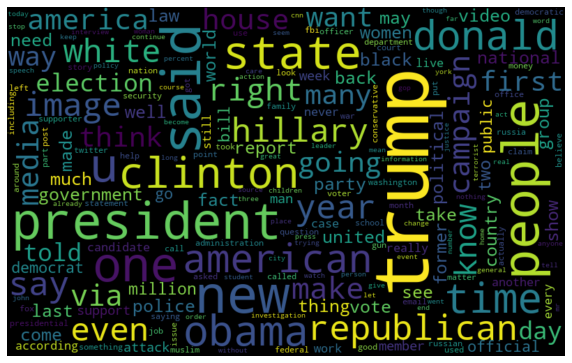
Before cleaning the data, a new column is created named ‘length\_before\_cleaning’ which shows the total length of the news respectively before cleaning the text.



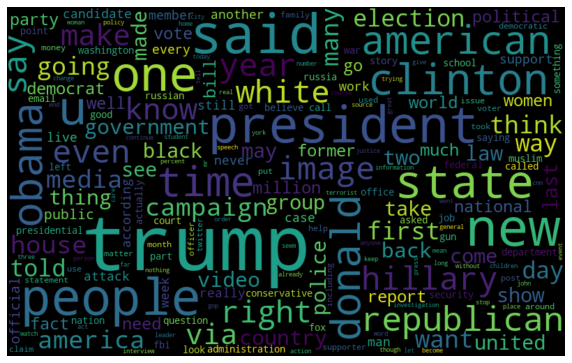
**Basic** **data** **exploration**



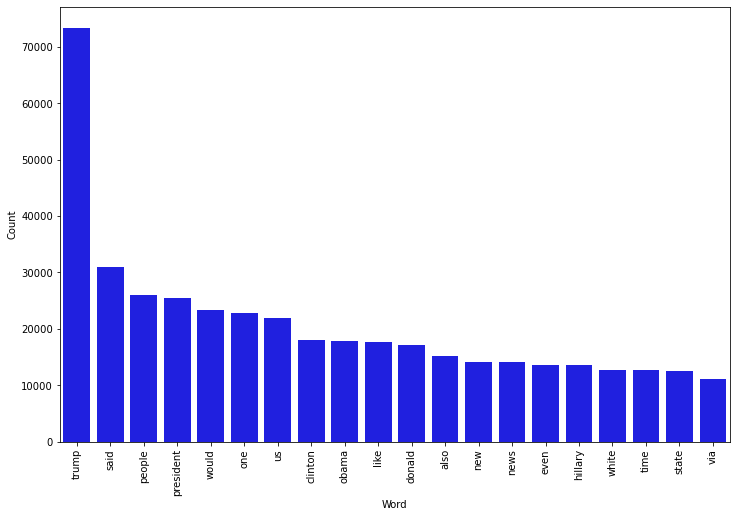


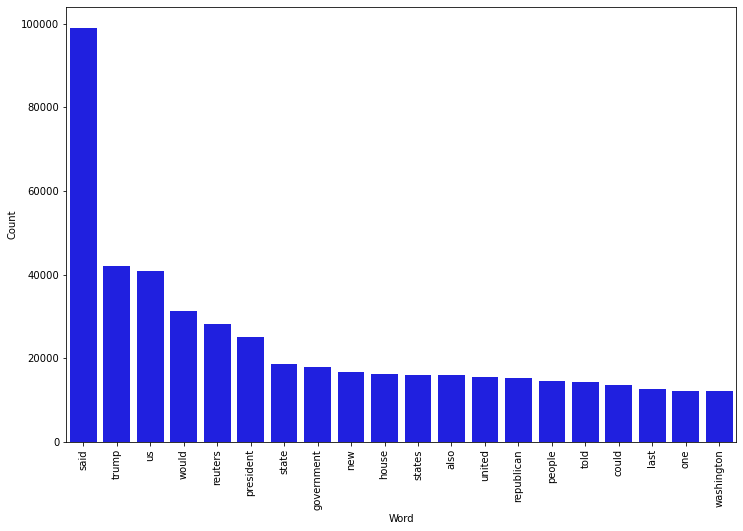






Most Frequent words in the fake News-



Most frequent words in real news

MODEL/S DEVELOPMENT AND EVALUATION

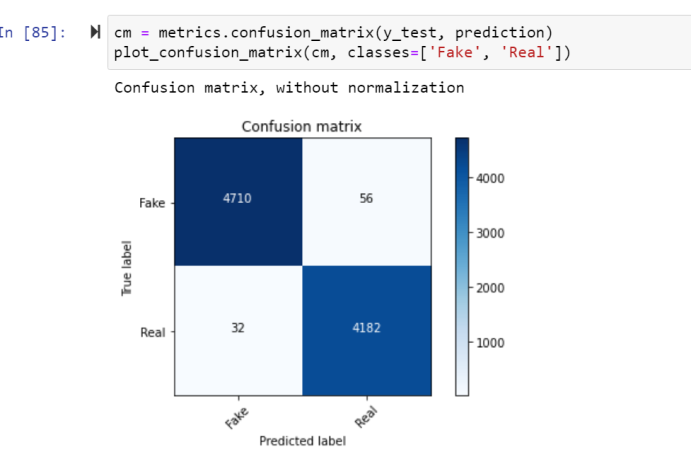
**Splitting** **the** **train** **and** **test** **data**

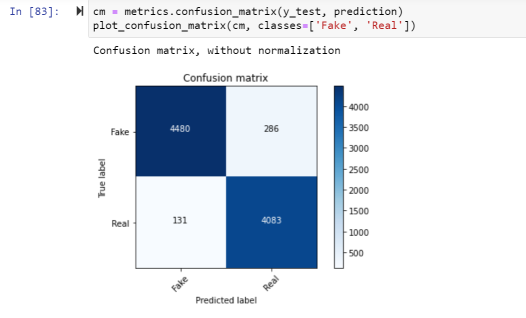
Training and testing the models minimize the effects of data discrepancies and better understand the characteristics of the model. The **training** data is used to make the machine recognize patterns in the data and the **test** data is used to see how well the machine can predict new answers based on its training. ‘X’ and ‘y’ were split for training and testing using train\_test\_split.

**Building** **the** **model**

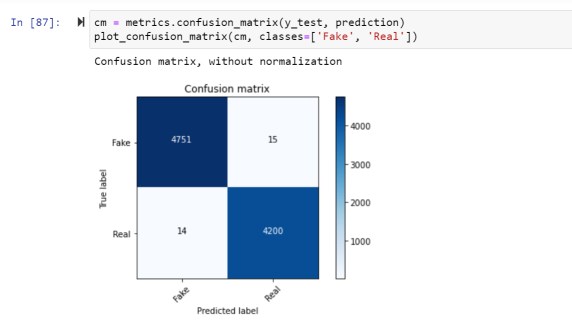
Running various algorithms, for the final results

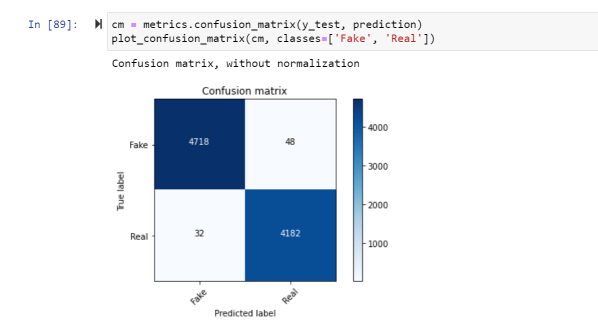
After running the algorithms and according to the scores of performance metrics and other scores, we can see that Logistic Regression and RandomForest Classifier algorithms are performing well. Now, we will perform Hyperparameter Tuning to find out the best parameters and try to increase the scores.

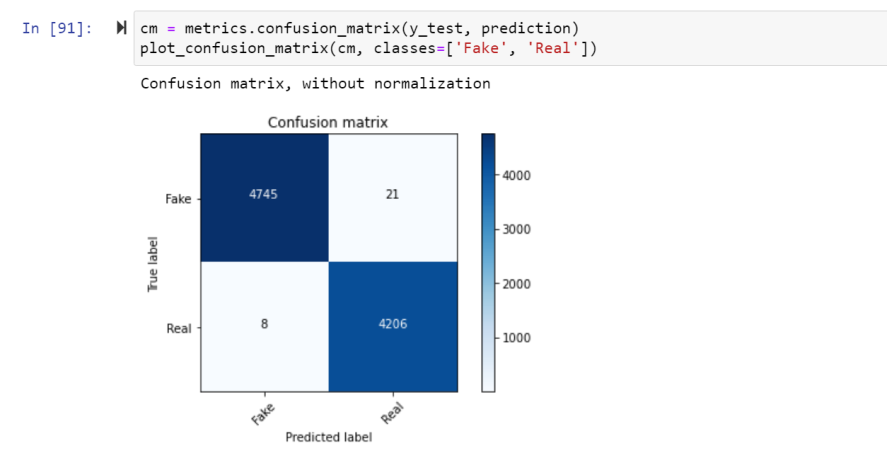


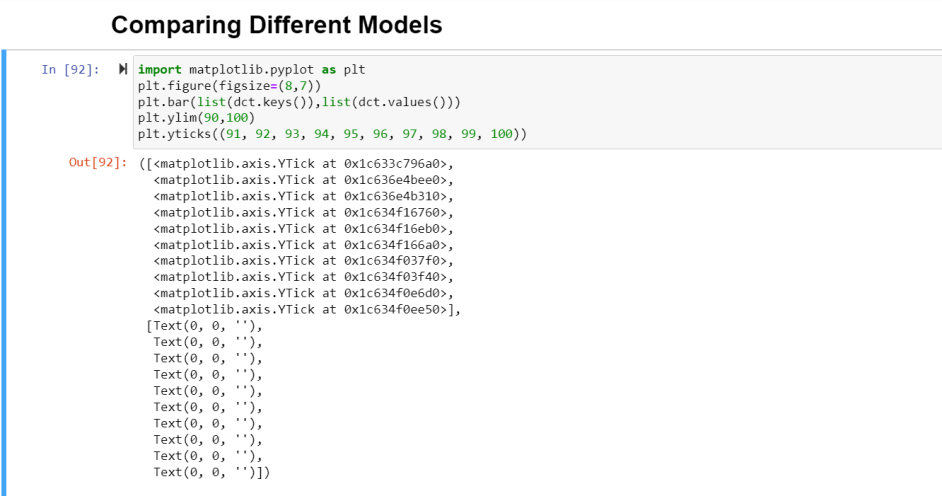


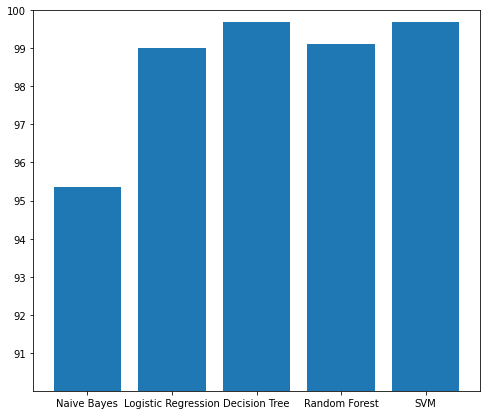












**CONCLUSION**

After the completion of this project, we got an insight of how to pre-process the data, analysing the data and building a model.

We did all the required pre-processing steps like checking null values, datatypes check, dropping unnecessary columns, etc.

-> We did the Exploratory Data Analysis using various plots and recorded the observations.

-> Using NLP, we pre-processed the comment text and did other steps like:

* Removing Punctuations and other special characters
* Splitting the comments into individual words
* Removing Stop Words
* Stemming and Lemmatising
* Applying Count Vectorizer
* Plotting wordcloud for knowing the weightage of words used

We found that LogisticRegression and RandomForestClassifier were performing well. The next step was to perform hyperparameter tuning

technique to these models for finding out the best parameters and trying to improve our scores.

LogisticRegression Algorithm gave us good scores and metric values than RandomForestClassifier and therefore we finalized it as the best model.