



**Submitted by:**

AVINASH SINGH

ACKNOWLEDGMENT

In preparation of this Project, I had to take the help and guidance of some video tutorials , article on medium and several google pages. As the completion of this Project gave me much pleasure. I would like to expand my gratitude to all those who have directly and indirectly who helped me to use wide thinking to explore here in dataset.

In addition, a thank you to my mentor Nitin Mishra, who introduced me to the Methodology of work, and who support always wherever I stuck ever. I also thank Flip Robo Technologies for providing such an opportunity to work on various types of projects which gradually improve my vision to apply for the datasets.

INTRODUCTION

**Business Problem Framing:**

The problem is not only hackers, going into accounts, and sending false information. The bigger problem here is what we call “Fake News”. A fake are those news stories that are false: the story itself is fabricated, with no verifiable facts, sources, or quotes. When someone (or something like a bot) impersonates someone or a reliable source to false spread information, that can also be considered as fake news. In most cases, the people creating this false information have an agenda, that can be political, economical or to change the behaviour or thought about a topic.

**Conceptual Background of the Domain Problem:**

The problem is real and hard to solve because the bots are getting better are tricking us. Is not simple to detect when the information is true or not all the time, so we need better systems that help us understand the patterns of fake news to improve our social media, communication and to prevent confusion in the world.

**Review of Literature:**

Whether or not you’re a Republican or Democrat, social-media junkie or forested hermit, you have to admit that our rapidly digitalizing, polarized world is running into a certain problem: **fake news**. Whether it’s Facebook and the whole spectrum of monolithic technology companies feeding us only information that we want to hear to increase advertising revenue, we are losing touch on diversified information, and as consumers, are facing the likelihood that whatever we are fed may not even be accurate.

**Motivation for the Problem Undertaken:**

For the coders and experts, I’ll explain the Python code to load, clean, and analyse data. Then we will do some machine learning models to perform a classification task (fake or not).

Analytical Problem Framing

**Mathematical/ Analytical Modelling of the Problem:**

The modelling process will consist of vectorising the corpus stored in the “text” column, then applying [TF-IDF](http://www.tfidf.com/), and finally a classification machine learning algorithm. Pretty standard in text analytics and NLP. Here we are dealing with two main text columns which held some importance of the data. I prefer to select on focus more on the words which has great value of importance in the context. TF-IDF is the NLP terms I am going to apply on both these columns. This converts the important words a proper vectors with some weights.

**Data Sources and their formats:**

I received the dataset for this project through my mentor on this project and is in the .csv file. Some of the info such as Unnamed: 0,id,written\_by are seems unnecessary with respect to the problem.



**Data Pre-processing:**

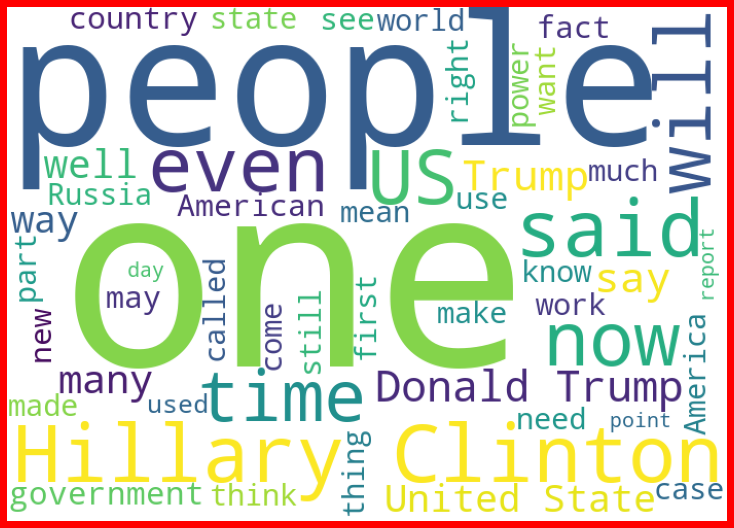
Removing the stop-words, punctuations and special characters from the text from both feature is done here.

**Visualization:**

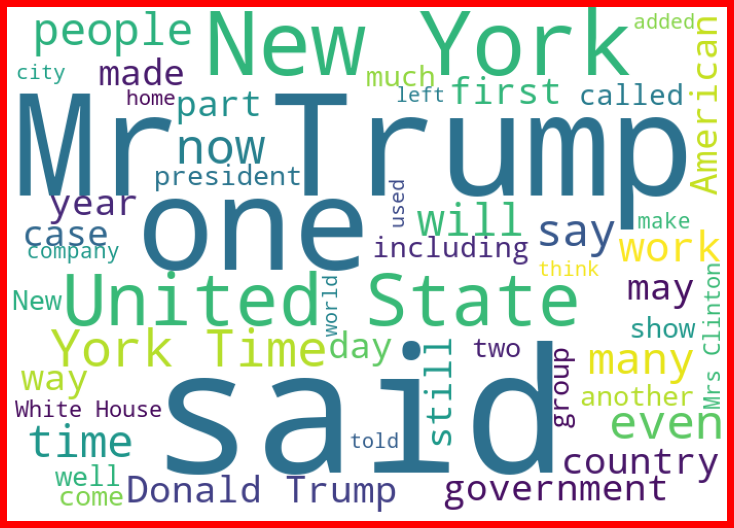
* How many fake and real articles?



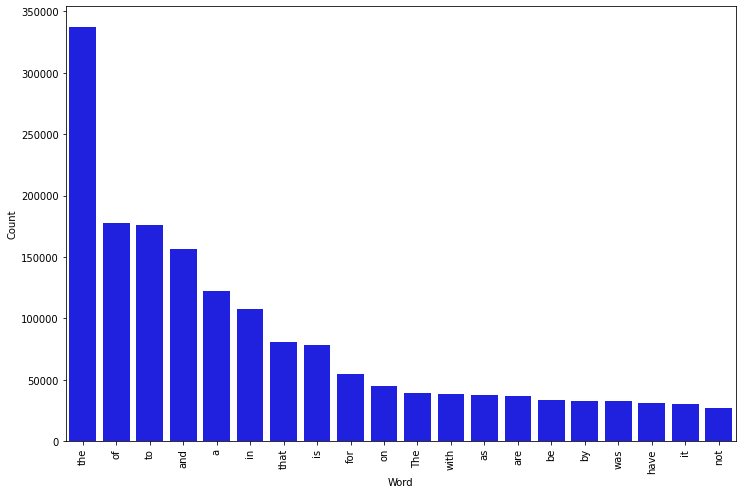
* Word Cloud for fake news:



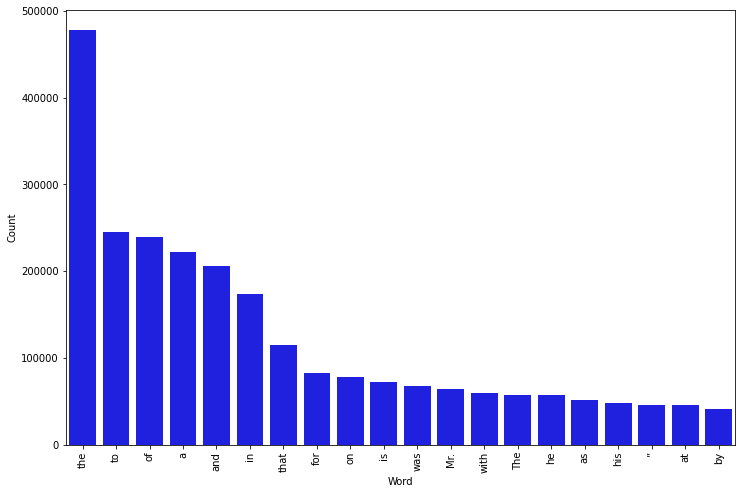
* Word cloud for real news:



* Most frequent words in fake news:



* Most frequent words in real news:



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

Text analytics and NLP can be used to work with the very important problem of fake news. The input data for the processing and getting the output is converted in to the vector form. It feed to the model in the form of series (one by one) which analyses by the model providing the certain score through the medium of performance metrics.

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

Hardware And Software required for this project:

Laptop with I3 processor and 8 GB Ram

Google colab

Python Pandas for processing

Scikit learn library

Model/s Development and Evaluation

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

To evaluate the performance of fake news detection algorithms, we used accuracy to evaluate classifiers in related areas: Accuracy = (𝑇𝑃 + )/(𝑇𝑃 + 𝑇𝑁 + 𝐹𝑃+𝐹𝑁) where TP, FP, TN, FN represent true positive, false positive, true negative and false negative, respectively. We compared the proposed framework to other machine learning algorithms. Analysing the data and use proper pre-processing techniques to extract out important words. Applying algorithms one by one to check the performance using evaluation metrics.

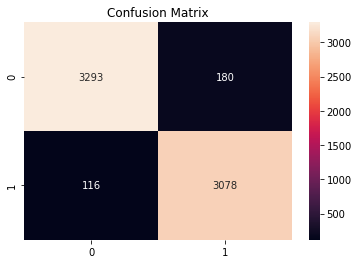
**Testing of Identified Approaches (Algorithms)**

Following are the algorithms thatapplied on this dataset. Random search is used for hyper parameter tuning the best model.

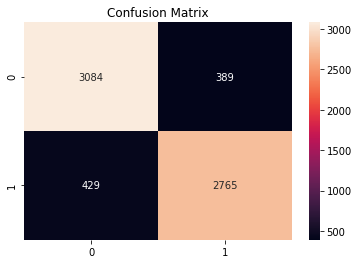
1. Logistic Regression
2. Gaussian NB
3. Decision Tree Classifier
4. Random Forest Classifier

**Run and Evaluate selected models:**

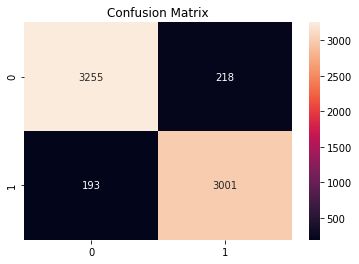
1. Logistic regression: Accuracy Score for LogisticRegression: 95.56%



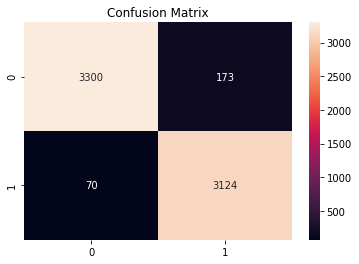
1. Gaussian NB**:** Accuracy Score for Naive: 87.73%

****

1. Decision Tree Classifier**:** Accuracy Score for DecisionTree: 93.8353%

****

1. Random Forest Classifier**:** Accuracy Score for Random Forest: 96.3552%

****

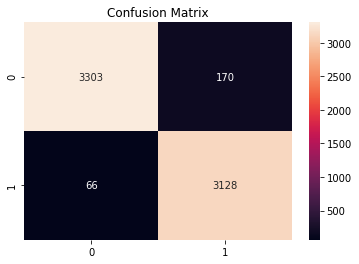
**Metrics for success in solving problem under consideration:**

Key metrics are F1-score, Precision, Recall and Roc-Auc Score.

Best model.

**Hyper parameter tuning for best estimators:**

Random forest is an ensemble learning method for text classification. From the above results it is observed that Random Forest is the best performing model. By comparing all metrics score, random forest is observed to be the best so it would be good to use for predicting fake news detection. The accuracy score of random forest is 96.35%. So hyper parameter tuning for best estimators that is random forest. I choose Random Forest Classifier as my final model because it give good accuracy score.

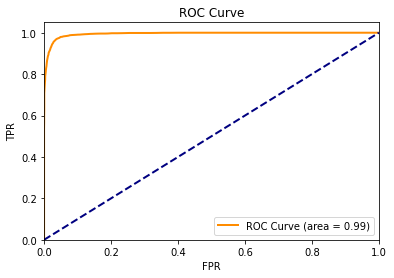
****

Accuracy = 0.9646017699115044

Precision = 0.9484536082474226

Recall = 0.9793362554790231

F1 Score = 0.9636475662353666

****

.

CONCLUSION

Text analytics and NLP can be used to work with the very important problem of fake news. We have seen the big impact they can have on people’s opinions, and the way the world thinks or sees a topic.

Within the vast domain of social media and its issues, there are critical problems which are a threat to the social media as a platform. Among those, fake news are not platform specific and more critical because of its effect over the drop in platform usage. We are trying to build a proactive methodology which could leverage the logistical information for better classification. Random Forest is the best performing model. By comparing all metrics score, random forest is observed to be the best so it would be good to used for predicting fake news detection. The accuracy score of random forest is 96.64%.From displaying the data, it seems there is null values present in the data. So, it is better to proceed by dropping it. As the above data is in text, so presence of special characters and stopwords is always there. After proper cleaning and processing, random forest classifier gives the highest accuracy as well as roc score.