

**Fake News Detection**

**Submitted by:**

**Hitesh Singh**

**ACKNOWLEDGMENT**

In the present world of competition there is a race of existence in which those are having will to come forward succeed. Project is like a bridge between theoretical and practical working. With this willing I joined this particular project.

I have taken efforts in this project. However it would not have been possible without the kind support and help of many individuals. I would like to extend my sincere thanks to all of them.

I would like to express my special thanks to my SME Tushar Sarashwat who gave me the golden opportunity to do this wonderful project on the topic **Fake News Detection**, which also helped me in doing a research and I came to know about so many new things.

I am really thankful to him.

I would also thankful to the online platforms who help me a lot in finishing this project within the limited time.

THANKS AGAIN TO ALL WHO HELPED ME.

**INTRODUCTION**

* **Business Problem Framing**

Fake news detection has recently attracted a growing interest from the general public and researchers as the circulation of missinformation online increases, particularly in media outlets such as social media feeds, news blogs, and online newspapers.For instance, a recent report by the Jumpshot

Tech Blog1 found that Facebook referrals accounted for 50% of the total traffic to fake news sites and 20% total traffic to reputable websites. Since the majority of U.S. adults –62%– gets news on social media (Jeffrey and Elisa, 2016), being able to identify fake content in online sources is a pressing need.

* **Conceptual Background of the Domain Problem**

So we build a machine learning model and and neural network that helps to understand that which news is fake and which news is real based on the sample data. We also do Exploratory Data Analysis to visualize the data graphically which is esay to understand.

* **Review of Literature**
* The proliferation of misleading information in everyday access media outlets such as social media feeds, news blogs, and online newspapers have made it challenging to identify trustworthy news sources, thus increasing the need for computational tools able to provide insights into the reliability of online content.
* In this project, we focus on the automatic identification of fake content in news. Our contribution is twofold. First, we introduce the datasets for the task of fake news detection, covering different news domains. We describe the collection and present several exploratory analysis on the identification of linguistic differences in fake and legitimate news content. Second, we conduct a set of learning experiments to build accurate fake news detectors.

* **Motivation for the Problem Undertaken**
* Every investigation begins with ideas that are further developed and inspired to address a variety of situations and circumstances.
* The client wants some predictions that could help them in further investment and improvement in selection of news. So to help them we make this project.
* My motivation behind this project is to do the proper research because research as a process for finding a solution to a problem after making a deep analysis and conducting studies of relevant factors. In general, research is a method designed to ensure that the information obtained is reasonable and supported by the quantitative and qualitative data, and that involves a systematic process. It includes the process of designing research methods, collecting and describing.

**Analytical Problem Framing**

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

1. First we check the basic information of the dataset that is it’s shape and it’s information using pandas library. The basic information tells the data type of our column and number of data present.
2. Then we check the unique information of our data columns.
3. After that we check for null values, if it present then we remove it because our data is text data and we cannot fill it.
4. To visualize the amount of missing values in different-2 columns we use Missingno library.
5. We have to also check whether our dataset is balanced or not.
6. We also create new columns to check the length of data before and after cleaning the news column to check the distribution of our data.
7. We use seaborn library to plot the target data and using wordcloud to for getting the sense of loud words in Real and Fake News.
8. Similarly we can also make wordcloud for single indexs where you can check the loud words for one index at a time.

* **Data Sources and their formats**

In this project the sample data is provided to us from our client database. The dataset is in csv (comma seprated values) format.

The data contains a CSV files train\_news.csv and we have to use the train\_news.csv data to build a model to predict whether a news is fake or not fake.

**Data- Description:**

There are 6 columns in the dataset provided to you. The description of each of the column is given below:

“id”: Unique id of each news article

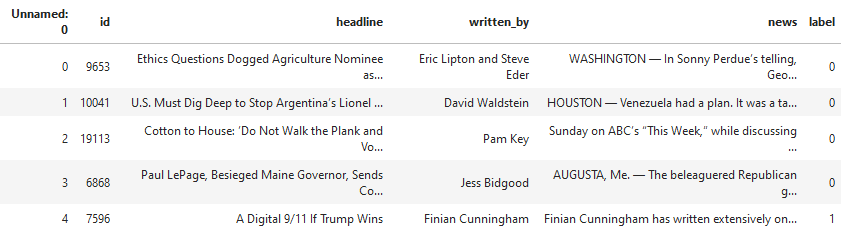
“headline”: It is the title of the news.

“news”: It contains the full text of the news article

“Unnamed:0”: It is a serial number

“written\_by”: It represents the author of the news article

“label”: It tells whether the news is fake (1) or not fake (0).



* **Data Preprocessing Done**
* In data cleaning we use mailny five steps using function:
* Removing HTML tags
* Removing special characters
* Converting everything to lowercase
* Removing stopwords
* Using Snowball Stemmer for Stemming
* *We create new column (clean\_length) after removing puncuations, stopwords from news column to check how much data is cleaned.*
* **Data Inputs- Logic- Output Relationships**

In this project, we focus on the automatic identification of fake content in news. Our contribution is twofold. First, we introduce the datasets for the task of fake news detection, covering different news domains. We describe the collection and present several exploratory analysis on the identification of linguistic differences in fake and legitimate news content. Second, we conduct a set of learning experiments to build accurate fake news detectors to see how the inputs column affects the target column.

* **State the set of assumptions (if any) related to the problem under consideration**
* After making the ML model we also make deep learning model using Neural networks because Neural network is not suffering from underfitting problem .
* In this we convert our text data into vectors via **Word Embedding.**
* Then we create our neural netwok with **LSTM** and **Bidirectional LSTM** both.
* For compilation of our model we use **Binary crossentropy** as our loss , **adam** as our optimizer and **accuracy** as our metrics.
* After training the model we check the performance metrics and accuracy.
* Whichever model gives the best result we save that model.
* **Hardware and Software Requirements and Tools Used**
* Hardware:
* Processor—Intel (R) Core(TM) i5-2430M CPU @ 2.40GHz
* Installed Memory(RAM)—8.00 GB
* System type—64-bit Operating System
* Software: Windos 10 Pro
* We have used Python Package because it is powerful and general purpose programming language.
* **NumPy**—It is a math library to work with ndimensional arrays. It enables us to do computation effectively and regurarly. For working with arrays, dictionary, functions data type we need to know NumPy.
* **Pandas**—It is high level Python library and easy to use for data importing , manipulation and data analysis.
* **Matplotlib**—It is a plotting that provide 2D and 3D plotting.
* **Seaborn**-- Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
* **SciPy**—It is a collection of numerical algorithm and domain specific tool boxes including optimization, statistics and much more.
* **Scikit-learn**—It is a collection of tools and algorithm for machine learning. It works with NumPy and SciPy and it is easy to implement machine learning models.
* **NLTK**-- NLTK is a leading platform for building Python programs to work with human language data.
* **Tensorflow**-- TensorFlow is an end-to-end open source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets researchers push the state-of-the-art in ML and developers easily build and deploy ML powered applications

**Model/s Development and Evaluation**

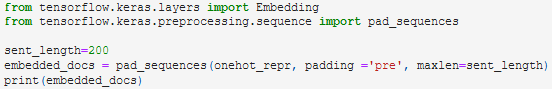
* **Identification of possible problem-solving approaches (methods)**
* Before making the model we convet our text into vectors so for that we use mainly two techniques:
* **TF-IDF Vectorizer**

****

* **Hashing Vectorizer**

****

* We uses **MultinomialNB** and **PassiveAggressive** Classifier to make the model with these two text conversion techniques.
* This is classification problem so we use accuracy score, classification report and confusion matrix as our eveluation matrix.
* As we know this dataset is balance although we evaluate our model on the precision and recall value along with f1\_score.
* For Neural network model we convert our text data into vectors via **Word Embedding**.

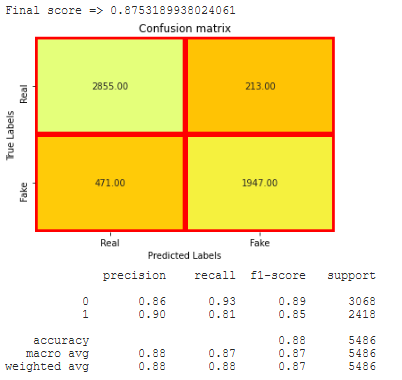


* For compilation of our Neural Network model we use **Binary crossentropy** as our loss , **adam** as our optimizer and **accuracy** as our metrics.
* **Testing of Identified Approaches(Algorithms)**
* We uses **MultinomialNB** and **PassiveAggressive** Classifier to make the ML models with these two text conversion techniques.
* We create our neural netwok with **LSTM** and **Bidirectional LSTM** both.
* For compilation of our Neural Network model we use **Binary crossentropy** as our loss , **adam** as our optimizer and **accuracy** as our metrics.
* **Run and Evaluate selected models**

**MultinomialNB using TF-IDF Vectorizer:**

****

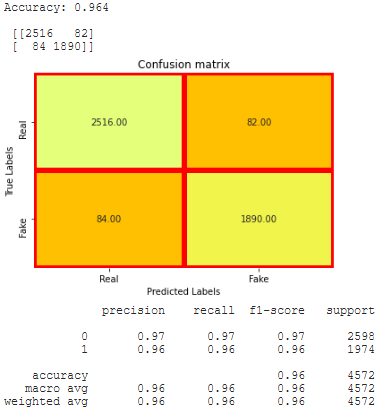
**Output:**

****

**Passive Aggrssive Classifier Using TF-IDF:**

****

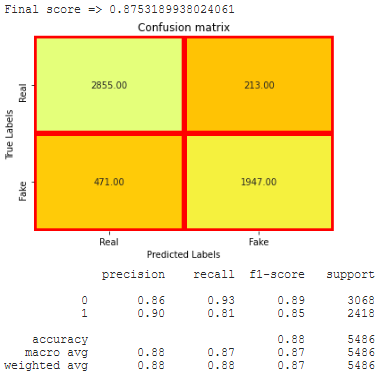
**Output:**

****

**MultinomialNB Using Hashing Vectorizer:**

****

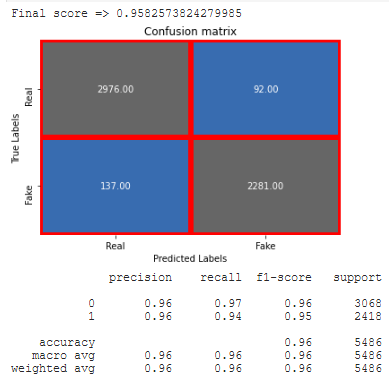
**Output:**

****

**Passive Aggressive Classifier Using Hashing Vectorizer:**

****

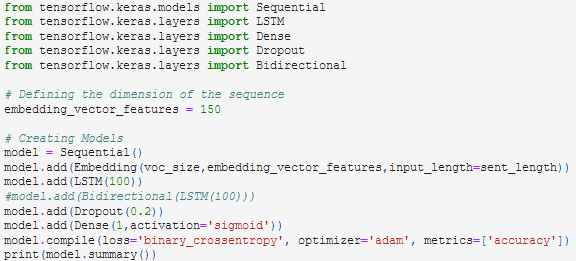
**Output:**

****

### Conclusion: In Between Multinomial NB and Passive Aggressive Classifier Algorithm using TFIDF and Hashing Vectorizer , Passive Aggressive Classifier with TFIDF gives the best result so we save this model as our final model..

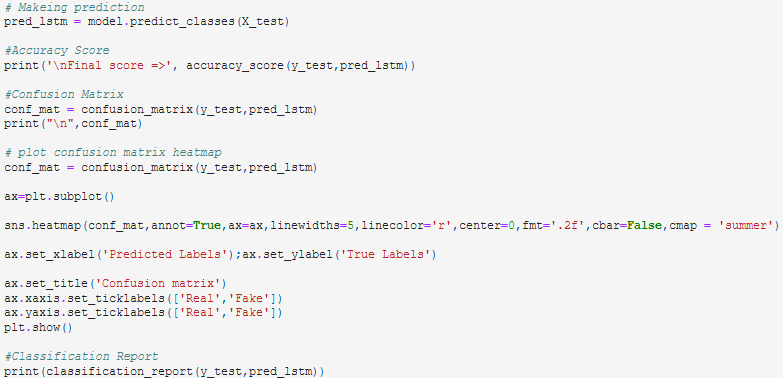
**Now we make the model via neural network see how its works**

**Using LSTM OR Bidirectional LSTM:**

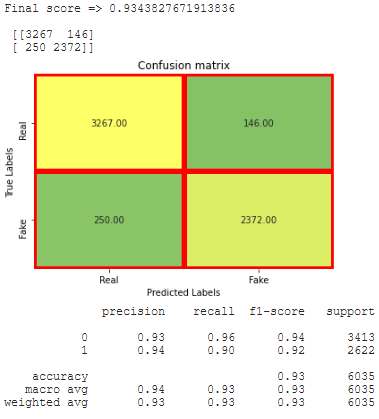
****

****

****

****

**Output:**

****

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

### Although it is a classification problem so we use accuracy score, classification report and confusion matrix as our evaluation metrics along with fi-score,precision and recall.

### As we know that our dataset is balance so we focus upon the accuracy score along with precision and recall value of our model.

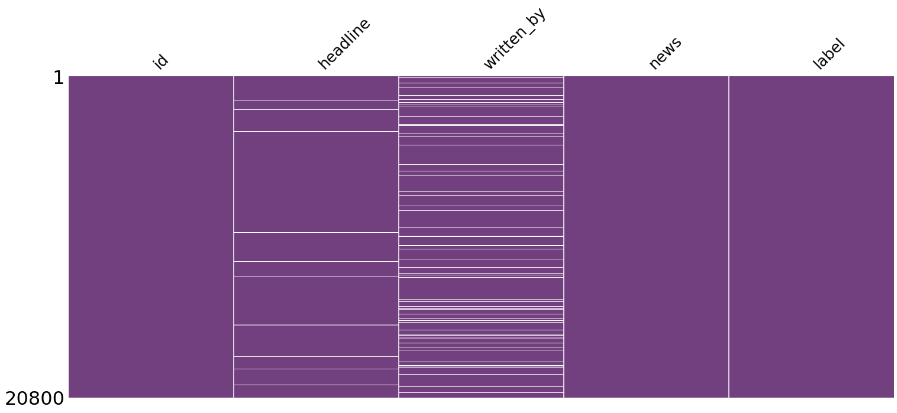
### Precision talks about all the correct predictions out of total positive predictions. Recall means how many individuals were classified correctly out of all the actual positive individuals.

For compilation of our neural network model we use **Binary crossentropy** as our loss , **adam** as our optimizer and **accuracy** as our metrics. After training the model we check the performance metrics and accuracy score along with classification report.

* **Visualizations**

It is the graphical representation of data that is used to check about the presence of outliers, patterns, distribution of the data, etc. There are different data visualisation libraries in python that include matplotlib, seaborn, etc. We will make use of the seaborn and matplotlib library to visualise the dataset.

**Plotting the Missing value matrix**

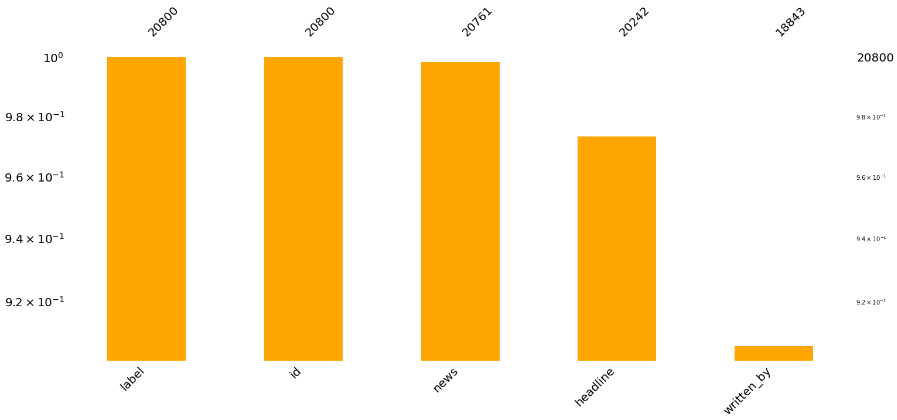
****

From this view it is clear that some of the columns have the missing values which needs to be filled with mean,median

or mode. But this dataset is text based so we cannot fill missing values with mean,median or mode. So we drop our

missing data.

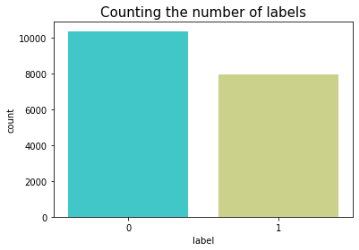
**Plotting the barplot of the null values:**

****

Missing number bar plot, plots the graph of the number of non-missing values as well as providing the actual number of

values present in the columns. We get an idea of how much of each column is missing.

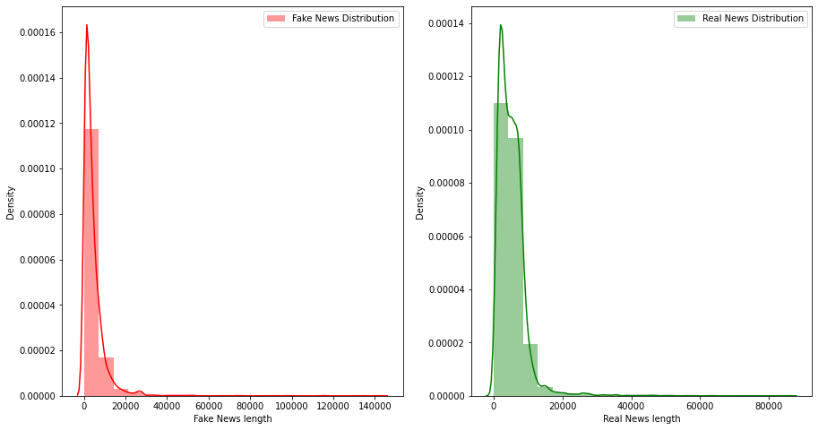
**Plotting the count of label:**

****

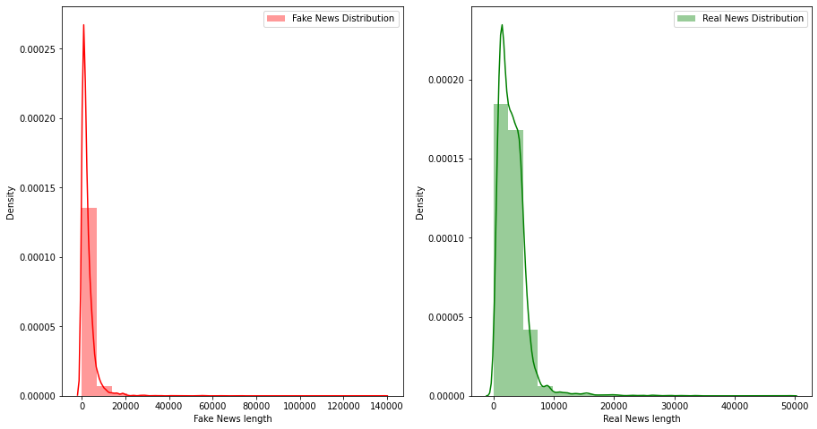
****

Plotting the distribution of Message before and after cleaning news:

Before cleaning:

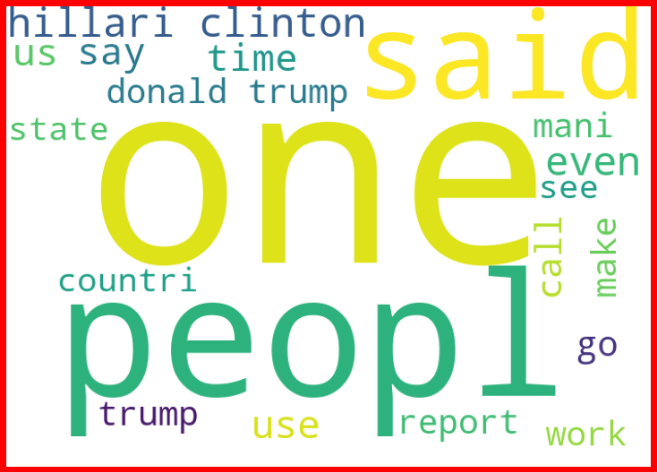


After Cleaning:

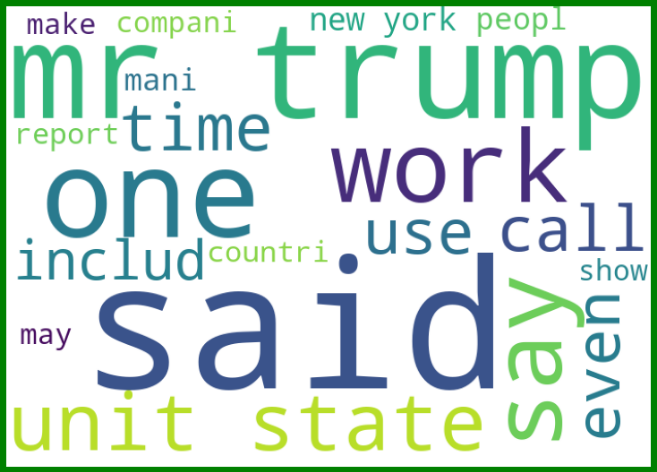


We can clearly see that after cleaning the data, the distribution of data reduces.

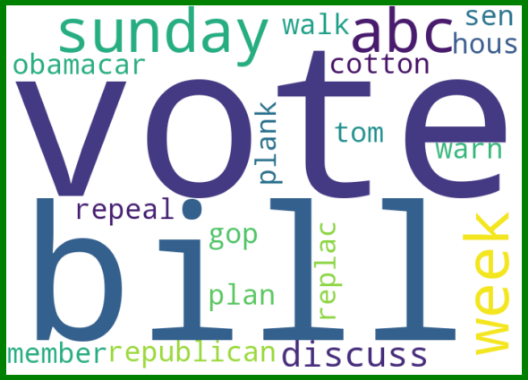
Getting sense of loud words in Fake News:



Getting sense of loud words in Real News:



Plotting the wordcloud for one index of news cloumn:



Similarly we can plot different-2 world cloud for different-2 indexs of news column.

* Interpretation of the Results

In this project we learn how to build a machine learning model for classification based problem.

We also learn how to make deep learning model using neural networks. Because neural networks handle higher data so effectively and model gives the better result.

The goal of any machine learning problem is to find a single model that will best predict our wanted outcome. Rather than making one model and hoping this model is the best/most accurate predictor we can make, ensemble methods take a myriad of models into account, and average those models to produce one final model.

On doing this project the biggest problem I have faced is that I am not able to use GridSearchCV. Because when I use GridSearchCV then my system takes too much time to give the result as our dataset is too large . So If I uses GridSearchCV then our result improves.

So based on all the learning and outcomes our Passive Aggressive Classifier with TF-IDF vectorizer gives the best result so we save this model as our final model by using Joblib as a pickle file.

## And in between LSTM and Bidirectional LSTM , LSTM model gives the best result so we save it as our final model in h5 format

**CONCLUSION**

* Key Findings and Conclusions of the Study
* There are null values present in our dataset so we drop it because we cannot fill the text data.



* The data set is balanced before removing the null values and after removing the null values it is still balanced.





* Headline column gives the headline of the news and there are 17931 unique headlines.
* News column gives the full news and there are 18017 unique news story.
* Written\_by column gives the name of writers and there are 3838 unique writers.
* There are 17931 headline and 18017 news articals. It seems that (18017-17931) = 86 news were duplicate or may be rewritten by some other writers.
* We using Missingno to plot the missing number matrix and barplot which makes easy to see the number of null values visually.
* Learning Outcomes of the Study in respect of Data Science
* In this project we learn how to build a machine learning model for classification based problem.
* We also learn how to make deep learning model using neural networks. Because neural networks handle higher data so effectively and model gives the better result.
* The goal of any machine learning problem is to find a single model that will best predict our wanted outcome. Rather than making one model and hoping this model is the best/most accurate predictor we can make, ensemble methods take a myriad of models into account, and average those models to produce one final model.
* On doing this project the biggest problem I have faced is that I am not able to use GridSearchCV. Because when I use GridSearchCV then my system takes too much time to give the result as our dataset is too large . So If I uses GridSearchCV then our result improves.
* So based on all the learning and outcomes our Passive Aggressive Classifier with TF-IDF vectorizer gives the best result so we save this model as our final model by using Joblib as a pickle file.

## And in between LSTM and Bidirectional LSTM , LSTM model gives the best result so we save it as our final model in h5 format

* Limitations of this work and Scope for Future Work

In this project the sample data is provided from our client database. In this project, we addressed the task of automatic identification of fake news. We introduced new fake news dataset, obtained through crowdsourcing and covering three news domains. We developed classification models as well as LSTM model that rely on a combination of lexical, syntactic, and semantic information, as well features representing text readability properties. Our best performing models achieved accuracies that are comparable to human ability to spot fake content.

As Today I uses i5 processor computer so I am not able to use GridSearchCV and cross validation because my system takes too much time to give the result. If some how I uses GridSearchCV and Cross validate the model then our model result will be surely improved.

I would conclude the project report by hoping that now you have understood every step that is required to be done to build a machine learning model. We have built the classification model for classifying the labels that the news which have label 1 that is Fake News and the news which have label 0 that is Real News and then evaluated it using different error metrics.