Categorization of News during pandemic

1) USE CASE

News is a crucial part which gives the information of current events. Due to the extensive use of social media, the roll out of fake news has been increased. Currently during the pandemic of COVID-19, a lot of fake news have created a panic in the public which can lead to a dissident in society. These news are created intentionally to misguide people and to bring down the authenticity. This has caused problems as it can change opinions and influence people. The main objective of this project is to use the efficient machine learning technique to train the model and verify the authenticity of the news.

2) DATA EXPLORATION

Datasets have been taken from Kaggle to train the model. It consists of 27865 data points. The links for above datasets are:

https://www.kaggle.com/c/fake-news/data

https://www.kaggle.com/jruvika/fake-news-detection

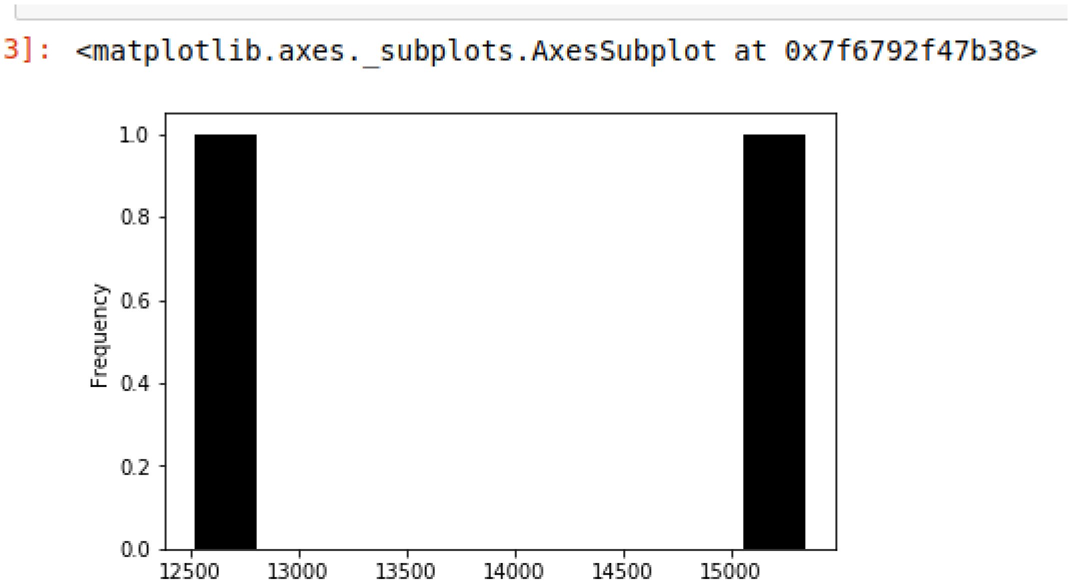
The features used to train the classifier are title and the content of news collected from the dataset.

3) DATA QUALITY ASSESSMENT

Scraped data from Kaggle had missing values, duplicates and NULL values. The quality of data is brought to standard by selecting the relevant columns from the dataset which could contribute to the efficient training and performance. Removal of the above inconsistencies has made the dataset perfect to work with.

4) DATA VISUALIZATION

After preprocessing the data, the number of fake and real news has been classified and plotted.



```
3]: <matplotlib.axes._subplots.AxesSubplot at 0x7f6792f47b38>
```

REAL	15343
FAKE	12522

5) FEATURE ENGINEERING

Data reduction, Data integration and Data cleaning are the important steps carried out in this process. The data has been converted to the required standards by removing duplicates, NaN, Null values and missing values. Total 27865 data points are obtained after data processing.

6) PLATFORM

IBM Watson is used as it is open source and helps us to build,train and deploy the applications. Jupyter Notebook with Python 3.6 , libraries like sklearn,keras,pandas,numpy,matplotlib,spacy,empath etc are used. Tensorflow is an open source software library which has been used in the backend of ML. NLP is used specifically in this project for converting the text to feature vectors. TF-IDF Vectorizer is the algorithm used for the above task.

7) ALGORITHM

After the news has been converted into feature vectors, the following algorithms are been implemented for the categorization:

1. **XGBoost Classifier**

   XGBoost is the open source implementation of gradient boosted decision trees which has been designed to improve speed and performance. The dataset contains a measurable quantity of data points and so it was good to use XGBoost as it speeds up training and classifies categorical values efficiently. As NLP part is also involved in training, XGBoost is engineered to prove efficient in computing time and memory resources.The accuracy obtained using this classifier is 84%

2. Random Forest Classifier

It is a part of ensemble learning where the random forest is used for classification and the group with more votes becomes model prediction.Uncorrelated trees come together as a committee to outperform as an individual model. The dataset suits this algorithm and gives the accuracy of about 87%.

3. **Neural Network (Deep Learning Model)**

   Deep Neural Network has been used which involves hyper-parameter tuning,regularization and optimization. In project, 1 Conv layer, Dropout layers for regularization,Max Pooling layer to average out max vectors, Adam optimizer, relu as activation. Sigmoid activation function is used after the dense neural networks to classify the fake and real news. The model is trained in 5 epochs with 22292 points trained in 1 epoch.

4. TFIDF Vectorizer (NLP)

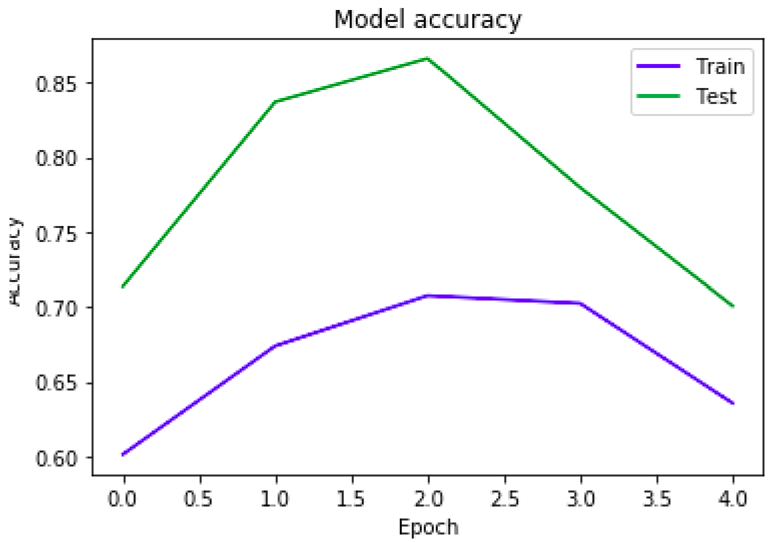
TF-IDF is a method used to represent text in a format which can be easily processed by the machine learning algorithms. The importance of a word is proportional to the number of times the word appears in the document but inversely proportional to the number of times the word appears in the corpus. The weights are composed of 2 terms:

1. TF (Term Frequency) - Calculates the frequency of word

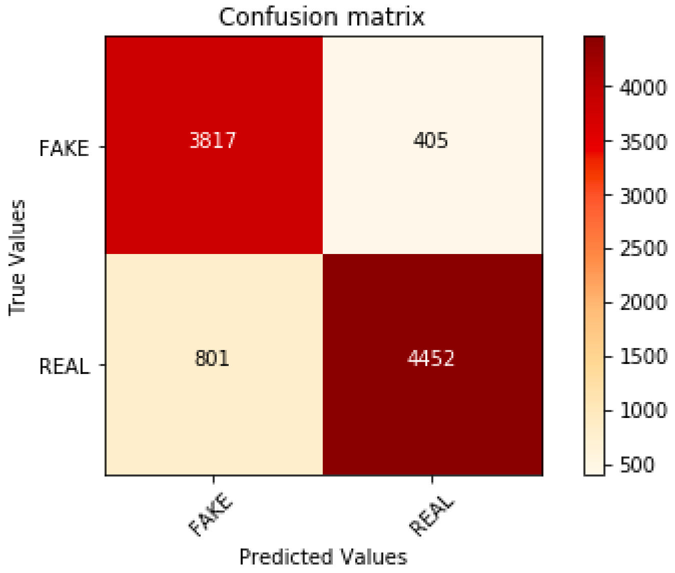
2. IDF (Inverse Document Frequency) - Denotes the importance of word in document.

8) MODEL PERFORMANCE INDICATOR

Model evaluation is done using Confusion Matrix, F1 score, Precision and Recall. Accuracy Indicator has been used to assess the model through training and testing data. The probabilities of True Positive and False Positive Rates have been visualized using a confusion matrix.



```
Confusion matrix
```



Confusion Matrix for XGBoost Classifier

9) DATA SOURCE

The dataset has been scraped from Kaggle. CSV format dataset has been used for the project and the standards are specified.

10) ENTERPRISE DATA

IBM Watson Studio has been used as a cloud based solution as it is open source and it can be used to extend the application. Everyday lots of data feeds and news are upcoming and added to the dataset. To deal with this ever increasing data, Watson can be used wherein subsets of data can be transferred and stored.

11) STREAMING ANALYTICS

Real time data can be collected from social media and news media. Data preprocessing can be done by collecting real time data and making the data structured for real time data processing.