FAKE NEWS DETECTION WEB APP

BY TEAM APACHE

Meet Team Apache

GROUP MEMBERS:

- ORINA TOLULOPE O.
- ABIOLA LAWANI
- YAHYA MARYAM ADEOLA
- AKINBO GBEKELEOLUWA
- SINJINI GHOSH
- OKPARA ESTHER
- EROMO OKONEDO
- JOHN OLUTOKI
- ATEHE STEPHEN
- DAVID OLOYEDE

PROBLEM STATEMENT

Our Task in this project is to build a Fake news Classifier; such that given a news content, the model would be able to predict if it is a True News or Fake News.

GOAL

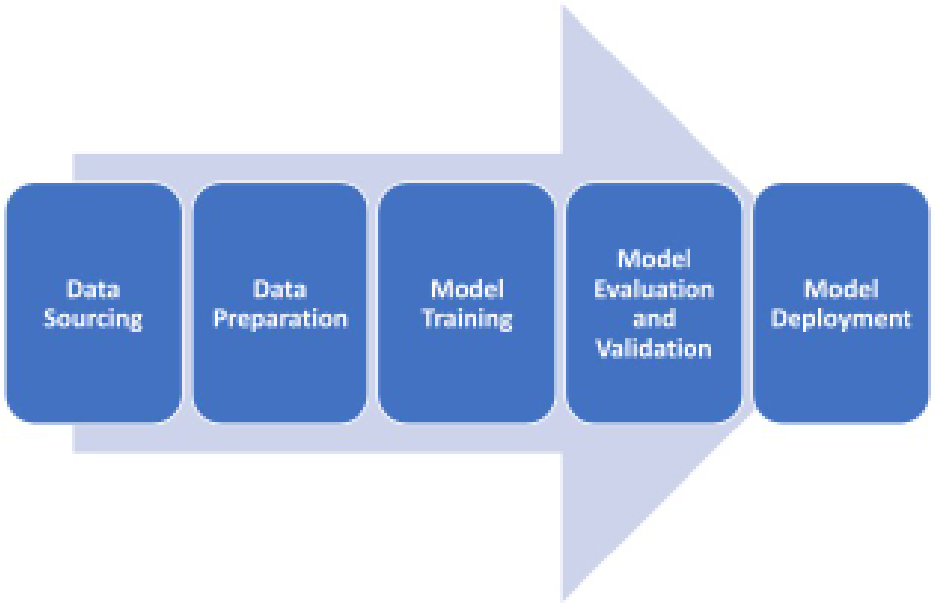
- The objective of this presentation is to outline end-to-end steps taking in building and training a machine learning model to classify fake and true news
- Using the best performing algorithm we deploy via Streamlit.

Project Scope And Boundary

- A Text Classification Task on Natural Language Processing (NLP).
- The news niche focuses on political news in the United States, The news article examined in the dataset is 2 years old.
- Kaggle fake news twitter dataset was used for this analysis.

Link: https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

PROJECT WORKFLOW



DATA PROCESSING

Libraries used for this project

- Pandas for data analysis numpy for numerical computation
- Matplotlib for visualization
- spacy for information extraction to perform such as (NER, POS tagging, dependency parsing, word vectors
- nltk for text preprocessing, converting text into numbers for the model
- Seaborn for visualization
- Textblob for text preprocessing, such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation
- WordCloud for Visualization, and pickle for exporting model.

Project Methodology

Preparing the dataset the dataset from Kaggle is provided in 2 CSV files which are already classified between true and fake news. The dataset was loaded using the pandas library however since it is textual data we carried out data cleaning, pre-processing, EDA and model-building oper

- This is the loaded dataframe

In [6]:

```
fake_news = pd.read_csv('Fake.csv')
fake_news.head()
```

Out[6]:

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017

In [7]:

```
fake_news.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23481 entries, 0 to 23480
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   title    23481 non-null  object
 1   text     23481 non-null  object
 2   subject  23481 non-null  object
 3   date     23481 non-null  object
dtypes: object(4)
memory usage: 733.9+ KB
```

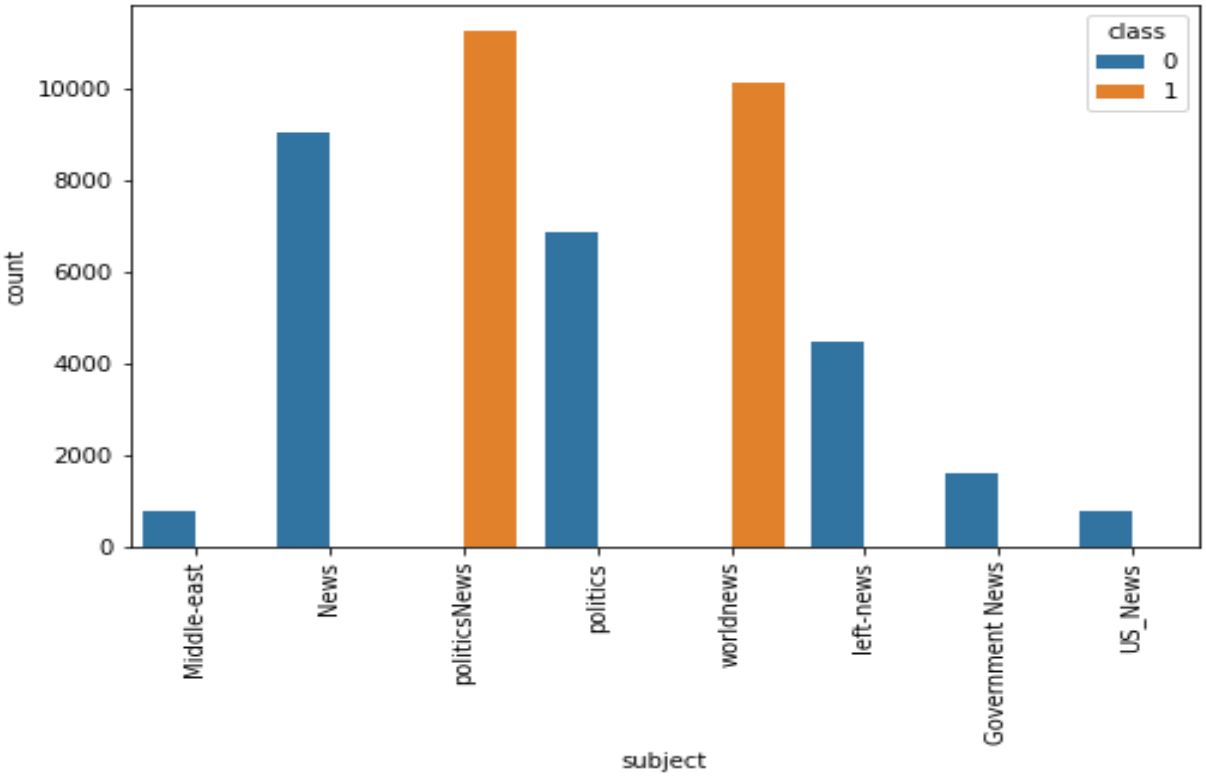
The dataset was examine for missing values, and it is interesting that there are no missing values we have a dataset of 4 features and 23481 observation

GENERATING THE WORDCLOUD FROM THE PREPROCESSED DATASET

WordCloud : is a visualization technique for text data wherein each word is picturized with its importance in the context or its frequency.

EXAMINING THE CLASS AND THE SUBJECT OF THE NEWS CONTENT IN THE DATASET

- True and Genuine News Articles are seen within the scope of political news and world news which can skew our model making it biased to a particular news domain.



PERFORMING NAME ENTITY RECOGNITION ON THE DATA

Named entity recognition (NER) – also called entity identification or entity extraction – is a natural language processing (NLP) technique that automatically identifies named entities in a text and classifies them into predefined categories.

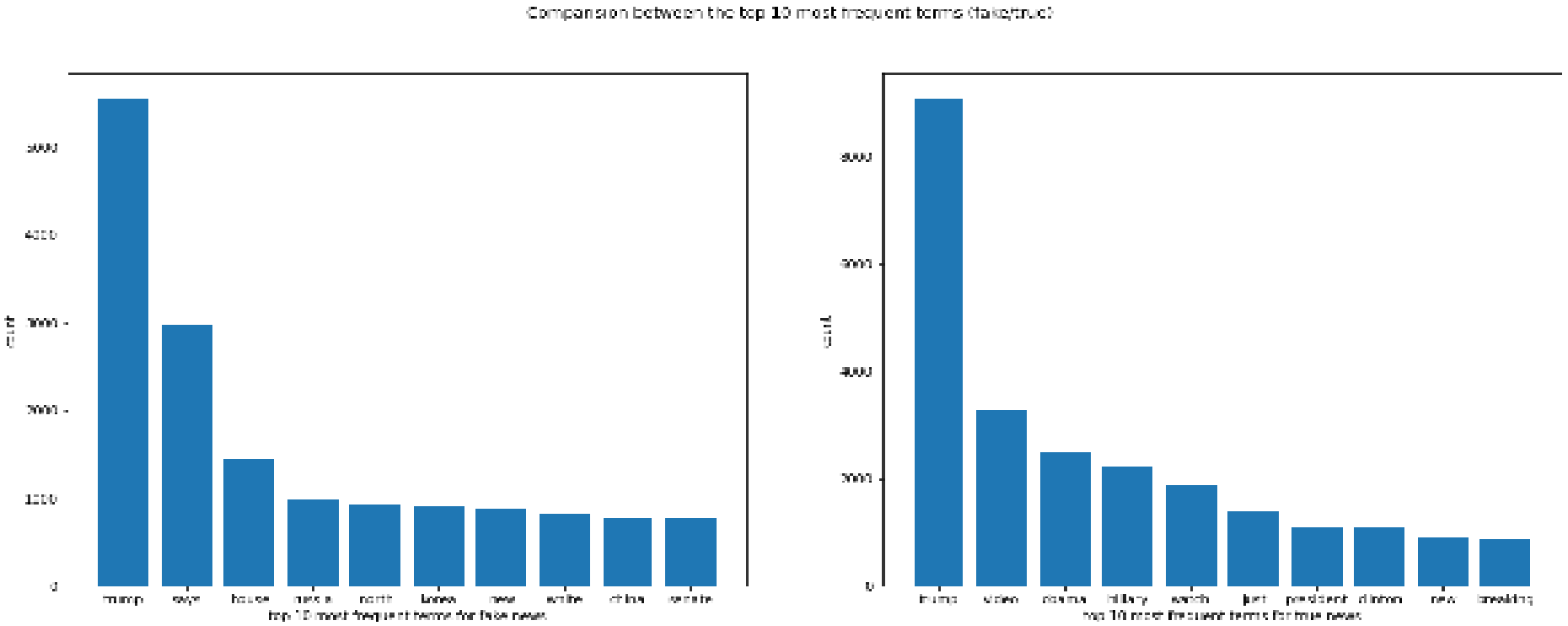
Name entity recognition was carried out on the text data to extract key names and entities present in the dataset and the below is the visualization

FEATURE ENGINEERING

In order to draw more insight from the dataset new features were engineered such as:

- Polarity : which is an output of the textblob which gives the ability of knowing the sentiment in each tweet.
- text_len : gives the length of each text or tweet size
- text_word_count : gives the count of word in the text
- title_len : gives the size of the tweet title

VISUALIZATION FOR THE MOST FREQUENT TITLE IN THE DATA



VISUALIZATION FOR THE MOST FREQUENT TITLE IN THE DATA CONTD.

- Observation from the above visualization:
- Based on the comparison between the top 10 frequent words in titles and news text, we can infer that both fake and true news is dominated by news relating to politics and more specifically, the subject being heavily related to American politics is shared between true and fake news. This would result in the model been biased to classifying news that relates to only American Politics and probably of that time frame. To mitigate this bias more recent data and diverse news data would be needed

MODEL SELECTION AND DATA PIPELINE

- **List of Classifier Models Use**

Classical machine learning algorithms were utilize for this classifier, a deep learning model was also used.

- Naive bayes - multinomial
- Logistics regression
- Random forest classifier
- Gradient boosting classifier
- LSTM layers featured with Word Embeddings - Deep learning model

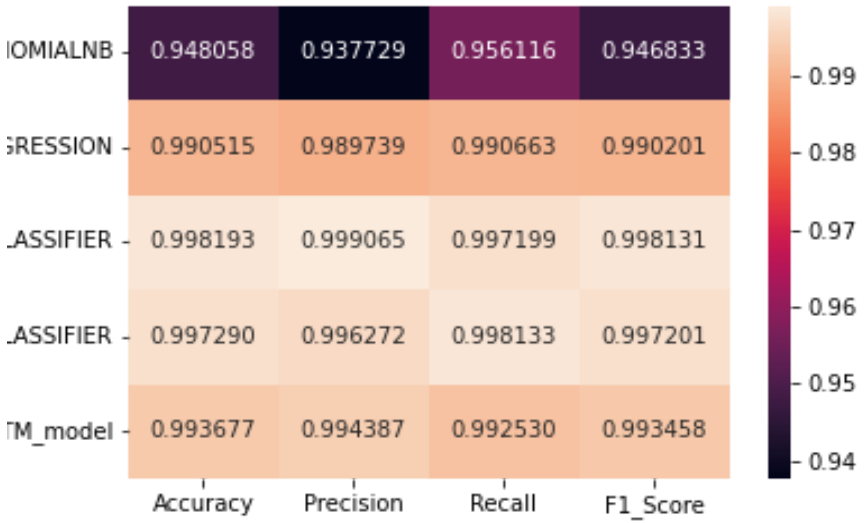
MODEL EVALUATION

The performance of the model on the validation is examine using the below evaluation metrics

- Confusion matrix
- Accuracy score
- Precision,
- Recall,
- f1score

MODEL EVALUATION CONTD.

- The models were evaluated and the result shown in the table below:



MODEL DEPLOYMENT

The Fake News Classification app is then deployed on the web using streamlit and readily available to end users for use

- Fake news Classifiaction source link : https://github.com/apache21/fake_news_detection (https://github.com/apache21/fake_news_detection)

CONCLUSION

- From Our Project, we discovered that one can readily build a model that is capable of detecting a fake news
- Also, Given more data and time we could get more news data, label them and run it through a classifier model.


THANK YOU
ANY QUESTIONS?