FAKE NEWS PROJECT

Submitted by:

MIENGANDHA SINHA

ACKNOWLEDGEMENT

The success of this project depends largely on the encouragement from many others. I would like to express my gratitude to the Company to give this project to me. In making of this project I hereby used to take help from the references which is given by the company as sample documentation and details related to project and professionals and SME mentor for continuously guiding and tremendously helping me a lot in the project and the other previous projects helped me and guided me in completion of the project.

INTRODUCTION

- **Fake news**

  This project is all about the how fake news creates a big-to-big problems in our age. Nowadays fake news spreading like water and people share this wrong information without verifying it. This can impact serious problem on our online as well as offline discourse. One can even go as far as saying that, to date, fake news poses a clear and present danger to western democracy and stability of the society. It has rows and columns for acquiring the data and describing the time period that represents it. There are two datasets one for fake news and one for true news. In fake news, there is 23481 news and in true news, 21417 news.

- **Conceptual Background of the Domain Problem**

  The project aims to detect fake news because from the view of media, is the ability to attract viewers to their websites is necessary to generate online advertising revenue. This same goes to impose certain ideas and is often achieved with political agendas also. It can even say that, fake news poses a clear and present danger to western democracy and stability of the society. It is implemented by using some natural language processing methods like machine learning data only works with numerical features so we have to convert text data into numerical columns. So for the pre-process the text, data cleaning by stemming, lemmatization, remove stopwords, special symbols and numbers. After cleaning the data to feed the text data into vectorizer which will convert the text data into numerical features. It all worked by using Natural Language Processing.

- **Review of Project**

  There are two datasets one for fake news and one for true news. In true news, there is 21417 news, and in for fake news, there is 23481 news. Only it has to do by inserting label column zero for fake news and one for true news. Both datasets are combined by using pandas built-in function.

```
:
1  # replacing the labels for integers, necessary for the loss function
2
3  df['Label'] = result_data['Label'].replace({'Fake':0, 'True':1})
4
5  print("Number of Fake Articles: ",len(df.loc[df['Label'] == 0]))
6  print("Number of True Articles: ",len(df.loc[df['Label'] == 1]))
7
8  df.head()
```

```
Number of Fake Articles:  23496
Number of True Articles:  21402
```

:

	text	Label
8364	Not a day goes by when a prominent figure on t...	0
4671	WASHINGTON (Reuters) - U.S. House of Represent...	0
23034	This week President Donald Trump followed thr...	0
18694	BERLIN (Reuters) - Chancellor Angela Merkel s ...	0
10877	WASHINGTON (Reuters) - The U.S. House of Repre...	0

- **Motivation for the Problem Undertaken**

Here, the datasets have total of fake news have 23481 entries with 4 rows and true news have 21417 entries with 4 rows, need to insert one column each which carries zero for fake news and one for true news. By doing model building there need to use NLP for text-pre-process cleaning text by stemming, lemmatization, remove stopwords, remove special symbols and numbers and text data into vectorizer which will convert the text data into numerical features after cleaning this will works.

ANALYTICAL PROBLEM FRAMING

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

  In this project, mathematical/analytical modelling are used. Checking the null values found that having no null values in the datasets, the type of data frame is in pandas, data frame info tells that object, by using the data visualization they are:

  - the data frame shape and its info:

```
1  display(fake_data.shape)
2  display(true_data.shape)
3
4  display(fake_data.isnull().sum())
5  display(true_data.isnull().sum())
```

```
(23481, 5)

(21417, 5)

title      0
text       0
subject    0
date       0
Label      0
dtype: int64

title      0
text       0
subject    0
date       0
Label      0
dtype: int64
```

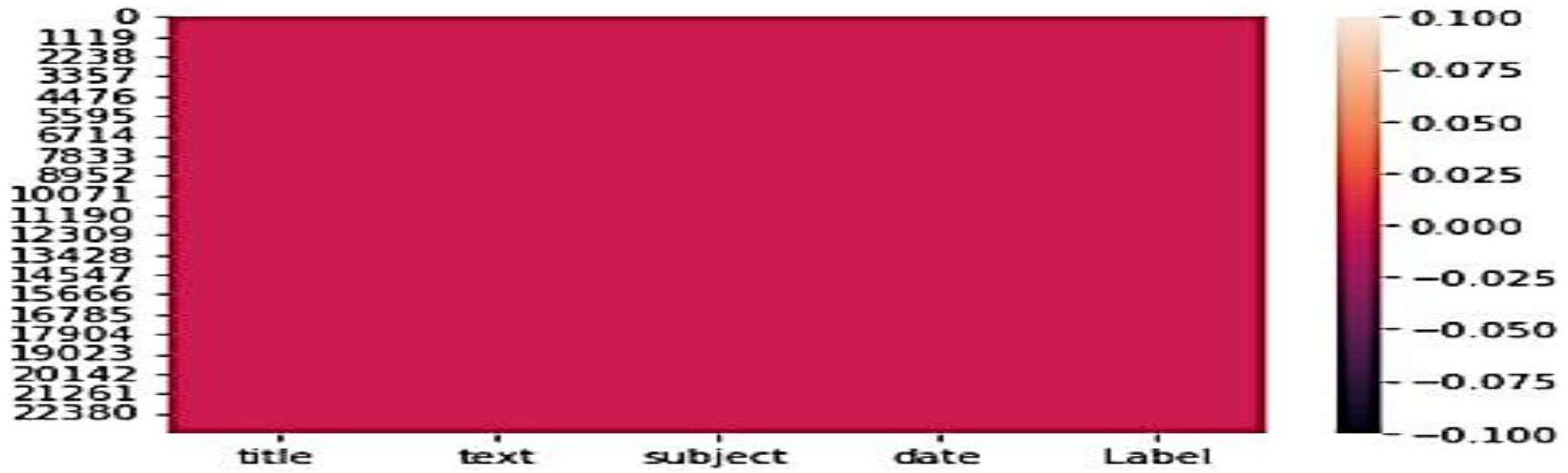
```
1  print(fake_data.info())
2
3  print(true_data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23481 entries, 0 to 23480
Data columns (total 5 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
 4   Label    23481 non-null  object
dtypes: object(5)
memory usage: 917.4+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21417 entries, 0 to 21416
Data columns (total 5 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   title    21417 non-null  object
 1   text     21417 non-null  object
 2   subject  21417 non-null  object
 3   date     21417 non-null  object
 4   Label    21417 non-null  object
dtypes: object(5)
memory usage: 836.7+ KB
None
```

- datasets:

```
2 sns.heatmap(fake_data.isnull())
```

```
<AxesSubplot:>
```




```
1 sns.heatmap(true_data.isnull())
```

```
<AxesSubplot:>
```

- **Data Sources and their formats**

  The data sources and their formats are from .csv file.

```
1 fake_data=pd.read_csv("Fake.csv")
2 fake_data.head()
```

	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

```
1 true_data=pd.read_csv("True.csv")
2 true_data.head()
```

	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017

- **Data Preprocessing Done**

  The steps followed for the cleaning of the data is Label Encoder after then importing preprocessing there transform the target columns into features then lastly it has to set for the data frame.

  There are some methods while doing data preprocessing:

```
1 data['Label'].value_counts()
```

```
Fake    23481
True    21417
Name: Label, dtype: int64
```

```
1 data['Label_encode'] = data['Label'].map({'Fake' : 0 , 'True' :1})
2 data.head()
```

	text	Label	Label_encode
0	Judge Jeanine: I wasn t even going to do an op...	Fake	0
1	For the third time in a row, Trump will head d...	Fake	0
2	As news that the CIA had solid evidence that R...	Fake	0
3	This new information just adds more validity T...	Fake	0
4	RIGA/TALLINN (Reuters) - Russia and its presid...	True	1

```
1 (round(result_data[result_data['Label']=="Fake"].shape[0]/result_data.shape[0],2))*100
```

```
52.0
```

```
1 result_data.skew()
```

```
Series([], dtype: float64)
```

Data cleaning

```
1 total =result_data.isnull().sum().sort_values(ascending=False)
2 percent = (result_data.isnull().sum()/result_data.isnull().count()).sort_values(ascending=False)
3 missing = pd.concat([total,percent], axis=1, keys=['Total' ,'Percent'])
4 missing.head()
```

	Total	Percent
title	0	0.0
text	0	0.0
subject	0	0.0
date	0	0.0
Label	0	0.0

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

  The relationships between inputs and outputs can be studied extracting weights of the trained model. Regression is that relationships between them can be blocky or highly structured based on the training data. It requires the data scientist to train the algorithm with both labeled inputs and desired outputs.

- **State the set of assumptions (if any) related to the problem under consideration**

  Presumptions are by using regression label encoding, classifier, selection of the best models, confusion matrix that it means the relationship between the dependent and independent variables look fairly linear. Thus, our linearity assumption is satisfied.

- **Hardware and Software Requirements and Tools Used**

  By importing many libraries are

```
# importing libraries
import numpy as np
import pandas as pd
import os
import spacy
import keras
import tensorflow as tf
from gensim.models import KeyedVectors
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
from sklearn.model_selection import train_test_split
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense,Embedding
from tensorflow.keras import layers

import warnings
warnings.filterwarnings('ignore')
```

MODEL/s DEVELOPMENT AND EVALUATION

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

Statistical analysis can be used in situations like gathering research interpretations, statistical modelling or designing surveys and studies. The collection and interpretation of data in order to uncover patterns and trends. It is a component of data analytics.

- **Testing of Identified Approaches(Algorithms)**

  There is no outliers.

- **Run and Evaluate selected models**

Building the model

```
inputs = keras.Input(shape = (vocab_size,), dtype="int64")
outputs = layers.Dense(1, activation="sigmoid")(inputs)
model = keras.Model(inputs, outputs)

model.compile(loss='mse',optimizer='SGD',metrics=['accuracy'])
epochs=30

model.summary()
```

```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 20000)]           0

 dense (Dense)               (None, 1)                 20001

=================================================================
Total params: 20,001
Trainable params: 20,001
Non-trainable params: 0
_________________________________________________________________
```

```
for index, topic in enumerate(nmf_model.components_):
    print(f'the top 15 words for topic #{index}' )
    print([tfidf.get_feature_names_out ()[i] for i in topic.argsort()[-15: ]])
    print('\n')
```

```
the top 15 words for topic #0
['supporters', 'said', 'white', 'featured', 'people', 'like', 'republican', 'image', 'just', 'realdonaldtrump', 'twitter', 'campaign', 'president', 'donald', 'trump']


the top 15 words for topic #1
['house', 'senate', 'america', 'republican', 'american', 'government', 'united', 'said', 'state', 'people', 'court', 'states', 'republicans', 'president', 'obama']


the top 15 words for topic #2
['spore', 'pst', 'hesher', 'alternate', 'episode', 'tune', 'animals', 'broadcast', '00', 'join', 'radio', 'room', 'pm', 'acr', 'boiler']


the top 15 words for topic #3
['party', 'candidate', 'presidential', 'secretary', 'election', 'state', 'bernie', 'democratic', 'email', 'emails', 'foundation', 'campaign', 'sanders', 'hillary', 'clinton']


the top 15 words for topic #4
['officials', 'house', 'james', 'security', 'flynn', 'putin', 'information', 'director', 'news', 'intelligence', 'investigation', 'russian', 'comey', 'russia', 'fbi']


the top 15 words for topic #5
['students', 'officer', 'twitter', 'school', 'man', 'com', 'lives', 'white', 'video', 'officers', 'said', 'gun', 'people', 'black', 'police']
```

- **Interpretation of the Results**

The results were interpreted from the preprocessing and modelling:

1. The predictions for news related politics, fashion and many more are not always correct.

2. So, it is clearly visible how much the quality and quantity of training data affects this fake news detection model.

3. If the model is trained with a more diverse dataset with news from various different domains, obtaining a much more robust and accurate classifier is not too far-fetched.

4. Aim to provide the user with the ability to classify the news as fake or real and also check the authenticity of the website publishing the news.

5. By using vectorizing data: TF-IDF that computes "relative frequency" that a word appears in a document compared to its frequency across all documents. TF-IDF weight represents the relative importance of a term in the document and entire corpus.

6. Accuracy was noted for all models.

7. Confusion Matrices for Static System.

CONCLUSION

Nowadays, the majority of the tasks are done online. Newspapers that were earlier preferred as hard-copies are now being substituted by applications like Facebook, Twitter, and news articles to be read online. Whatsapp's forwards are also a major source. The growing problem of fake news only makes things more complicated and tries to change or hamper the opinion and attitude of people towards use of digital technology. When a person is deceived by the real news two possible things happen- People start believing that their perceptions about a particular topic are true as assumed. Thus, in order to curb the phenomenon, have developed input from the user and classify it to be true news or fake news. To implement this, various NLP and Machine Learning Techniques have to be used. The model is trained using an appropriate dataset and performance evaluation is also done using various performance measures. The best model, i.e. the model with highest accuracy is used to classify the news headlines or articles, that it will be classified to its true nature.

THANK YOU