FAKE NEWS DETECTION USING NLP

TEAM MEMBER

NAME: J.JAYASREE

REG NO: 953421106019/au953421106019

Phase 3-Development Part 1

PROJECT: Fake News Detection

OBJECTIVE:-

The objective of fake news detection using Natural Language Processing (NLP) is to develop a system that can effectively differentiate between reliable and unreliable information in textual content.

Phase 3: Development Part 1

Introduction:

In this part you will begin building your project by loading and preprocessing the dataset.Begin building the fake news detection model by loading and preprocessing the dataset. Load the fake news dataset and preprocess the textual data.

USED DATA SET:-

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

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

Importing Libraries and Datasets:-

```
import pandas as pd

Import numpy as np

From sklearn.model_selection import train_test_split

From sklearn.feature_extraction.text import TfidfVectorizer,
CountVectorizer

Import matplotlib.pyplot as plt

Import itertools

From sklearn import svm

From sklearn.naive_bayes import MultinomialNB

From sklearn.ensemble import RandomForestClassifier,
GradientBoostingClassifier

From sklearn import metrics

Import spacy

From sklearn.feature_extraction.stop_words import
ENGLISH_STOP_WORDS

Import string

Import re

Import nltk

Import collections

From nltk.corpus import stopwords
```

```
From sklearn.feature_extraction import DictVectorizer
From sklearn.pipeline import Pipeline, FeatureUnion
From empath import Empath
From keras.preprocessing.text import Tokenizer
From keras.preprocessing.sequence import pad_sequences
Import pickle

df = pd.read_csv('Dataset/data.csv')
df.loc[df['Label']== 0, 'Label'] = 'REAL'
df.loc[df['Label']== 1, 'Label'] = 'FAKE'
df.columns
df['Label'].value_counts()
```

Out:

```
REAL    2137
FAKE    1872
Name: Label, dtype: int64
```

```
#Dropping the column URLs from the table
df.drop(['URLs'], axis = 1, inplace = True)
df.columns
```

Out:

```
Index(['Headline', 'Body', 'Label'], dtype='object')
```

```
#Selecting only fake news from all the types of news and then replacing the 'fake' by 0
df1 = pd.read_csv('Dataset/fake.csv')
df1.columns
df1['type'].value_counts()
df1 = df1.loc[df1['type']=='fake']
df1.loc[df1['type']== 'fake', 'type'] = 'FAKE'

#Selecting some columns from the table and renaming them\n",
df1 = df1[['title','text','type']]
df1.columns = ['Headline', 'Body', 'Label']
df1['Label'].value_counts()
```

Out:

```
FAKE    19
Name: Label, dtype: int64
df2= pd.read_csv('Dataset/fake_or_real_news.csv')
```

```
df2.columns
```

Out:

```
Index(['Unnamed: 0', 'title', 'text', 'label'],
dtype='object')
```

```
#Selecting few columns from the table and renaming the columns
df2 = df2[['title','text','label']]
df2.columns = ['Headline', 'Body', 'Label']
df2.columns
df2['Label'].value_counts()
```

Out:

```
REAL    3171
FAKE    3164
Name: Label, dtype: int64
```

```
df3 = pd.read_csv('Dataset/train.csv')
df3.columns
```

Out:

```
Index(['id', 'title', 'author', 'text', 'label'], dtype='object')
```

#Selecting few columns from the table and renaming the columns

```
df3 = df3[['title','text','label']]
df3.columns = ['Headline', 'Body', 'Label']
df3.loc[df3['Label']== 0, 'Label'] = 'REAL'
df3.loc[df3['Label']== 1, 'Label'] = 'FAKE'
df3.columns
df3['Label'].value_counts()
```

Out:

```
FAKE    10413
REAL    10387
Name: Label, dtype: int64
```

#Appending df1,df2,df3 to df

```
df = df.append(df1, ignore_index = True)
df = df.append(df2, ignore_index = True)
```

```
df = df.append(df3, ignore_index = True)


df = df.drop_duplicates()
# df.iloc[3647]
# print(df['Headline'][3647])
# print(len(df['Body'][3647]))
#df = df.dropna(how='any',axis=0)
cnt = 0
ind = []
for art in df['Body']:
    #print(type(art))
    if len(str(art)) < 10:
        ind.append(cnt)
    cnt+=1
df = df.drop(df.index[ind])
df
# print(df['Headline'][3647])
# print(len(df['Body'][3647]))
```

In [3]:

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
import matplotlib.pyplot as plt
import itertools
from sklearn import svm
from sklearn.naive_bayes import MultinomialNB
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
from sklearn import metrics
import spacy
from sklearn.feature_extraction.stop_words import ENGLISH_STOP_WORDS
import string
import re
import nltk
import collections
from nltk.corpus import stopwords
from sklearn.feature_extraction import DictVectorizer
from sklearn.pipeline import Pipeline, FeatureUnion
from empath import Empath
from keras.preprocessing.text import Tokenizer
from keras.preprocessing.sequence import pad_sequences
import pickle
```

```
Using TensorFlow backend.
```

In [4]:

```
df = pd.read_csv('Dataset/data.csv')
df.loc[df['Label']== 0, 'Label'] = 'REAL'
df.loc[df['Label']== 1, 'Label'] = 'FAKE'
df.columns
df['Label'].value_counts()
```

Out[4]:

```
REAL    2137
FAKE    1872
Name: Label, dtype: int64
```

In [5]:

```
#Dropping the column URLs from the table
df.drop(['URLs'], axis = 1, inplace = True)
df.columns
```

Out[5]:

```
Index(['Headline', 'Body', 'Label'], dtype='object')
```

In [6]:

```
#Selecting only fake news from all the types of news and then replacing the 'fake' by 0
df1 = pd.read_csv('Dataset/fake.csv')
df1.columns
df1['type'].value_counts()
df1 = df1.loc[df1['type']=='fake']
df1.loc[df1['type']== 'fake', 'type'] = 'FAKE'
```

In [7]:

```
#Selecting some columns from the table and renaming them\n",
df1 = df1[['title','text','type']]
df1.columns = ['Headline', 'Body', 'Label']
df1['Label'].value_counts()
```

Out[7]:

```
FAKE    19
Name: Label, dtype: int64
```

In [8]:

```
df2 = pd.read_csv('Dataset/fake_or_real_news.csv')
df2.columns
```

Out[8]:

```
Index(['Unnamed: 0', 'title', 'text', 'label'], dtype='object')
```

In [9]:

```
#Selecting few columns from the table and renaming the columns
df2 = df2[['title','text','label']]
df2.columns = ['Headline', 'Body', 'Label']
df2.columns
df2['Label'].value_counts()
```

Out[9]:

```
REAL    3171
FAKE    3164
Name: Label, dtype: int64
```

In [10]:

```
df3 = pd.read_csv('Dataset/train.csv')
df3.columns
```

Out[10]:

```
Index(['id', 'title', 'author', 'text', 'label'], dtype='object')
```

In [11]:

```
#Selecting few columns from the table and renaming the columns
df3 = df3[['title','text','label']]
df3.columns = ['Headline', 'Body', 'Label']
df3.loc[df3['Label']== 0, 'Label'] = 'REAL'
df3.loc[df3['Label']== 1, 'Label'] = 'FAKE'
df3.columns
df3['Label'].value_counts()
```

Out[11]:

```
FAKE    10413
REAL    10387
Name: Label, dtype: int64
```

In [12]:

```
#Appending df1,df2,df3 to df
df = df.append(df1, ignore_index = True)
df = df.append(df2, ignore_index = True)
df = df.append(df3, ignore_index = True)
```

In [13]:

```
df = df.drop_duplicates()

# df.iloc[3647]
# print(df['Headline'][3647])
# print(len(df['Body'][3647]))
#df = df.dropna(how='any',axis=0)
cnt = 0
ind = []
for art in df['Body']:
    #print(type(art))
    if len(str(art)) < 10:
        ind.append(cnt)
```

```
    cnt+=1
df = df.drop(df.index[ind])


df
# print(df['Headline'][3647])
# print(len(df['Body'][3647]))
```

Out[13]:

	Headline	Body	Label
0	Four ways Bob Corker skewered Donald Trump	Image copyright Getty Images\nOn Sunday mornin...	FAKE
1	Linklater's war veteran comedy speaks to moder...	LONDON (Reuters) - “Last Flag Flying”, a comed...	FAKE
2	Trump’s Fight With Corker Jeopardizes His Legi...	The feud broke into public view last week when...	FAKE
3	Egypt's Cheiron wins tie-up with Pemex for Mex...	MEXICO CITY (Reuters) - Egypt’s Cheiron Holdin...	FAKE
4	Jason Aldean opens 'SNL' with Vegas tribute	Country singer Jason Aldean, who was performin...	FAKE
5	JetNation FanDuel League; Week 4	JetNation FanDuel League; Week 4\n% of readers...	REAL
6	Kansas Tried a Tax Plan Similar to Trump’s. It...	In 2012, Kansas lawmakers, led by Gov. Sam Bro...	FAKE
7	India RBI chief: growth important, but not at ...	The Reserve Bank of India (RBI) Governor Urjit...	FAKE
8	EPA chief to sign rule on Clean Power Plan exi...	Scott Pruitt, Administrator of the U.S. Enviro...	FAKE
9	Talks on sale of Air Berlin planes to easyJet ...	FILE PHOTO - An Air Berlin sign is seen at an ...	FAKE
10	U.S. President Donald Trump Quietly Signs Law ...	By Aaron Kesel\nAs former White House chief of...	REAL
11	2017 Fantasy Football Team Defense Rankings - ...	2017 Fantasy Football Team Defense Rankings – ...	REAL
12	Just Shut Up & Play Some Damn Baseball!!	Just Shut Up & Play Some Damn Baseball!!\n(Bef...	REAL
13	Deloitte cyber attack affected up to 350 clien...	FILE PHOTO: The Deloitte Company logo is seen ...	FAKE
14	10/7: Chuck Axed; HBD Brickyard, Adam, Moonlig...	A Potato Battery Can Light up a Room for Over ...	REAL
15	Gunman’s Girlfriend Said She Didn’t Know He Pl...	• The authorities found evidence that the gunm...	FAKE
16	Marilou Danley, Gunman’s Girlfriend, Says She ...	The statement, which was read by her lawyer, M...	FAKE
17	Trump’s Immigration Rhetoric Echoes a Bitter F...	In bold documentary style, Retro Report looks ...	FAKE
18	Trump Bemoans ‘Little Appreciation’ As San Jua...	Red Flag Warning: These California Wildfires A...	REAL
19	In Meeting With Military, Trump Talks Of 'Calm...	In Meeting With Military, Trump Talks Of 'Calm...	REAL
20	Teacher Sparks Outrage By Asking Kids To Make ...	Red Flag Warning: These California Wildfires A...	REAL
21	9/28 Through the 40s: The Gloaming; HBD Bill, ...	Vietnam Is in Great Danger, You Must Publish a...	REAL
22	Weinstein Co board ousts Harvey Weinstein afte...	(Reuters) - The Weinstein Co has fired co-Chai...	FAKE
23	Hillary Clinton Suggests That Trump May Order ...	Hillary Clinton Suggests That Trump May Order ...	REAL
24	9/29 Through the 40s: HBD Cannonball & Paul, C...	Red Flag Warning: These California Wildfires A...	REAL
25	Sharapova storms into Tianjin quarter-finals	(Reuters) - Maria Sharapova overpowered Poland...	FAKE
26	10/3 Expo Park-Forbes Field Era: Pirates, Gray...	Red Flag Warning: These California Wildfires A...	REAL
27	Weinstein scandal no surprise to Hollywood	Chat with us in Facebook Messenger. Find out w...	FAKE
28	Blackhawks Roster Breakdown: Goalies	Blackhawks Roster Breakdown: Goalies\n(Before ...	REAL
29	With Christian Pulisic Driving, United States ...	When Pulisic tore open the left side of the Pa...	FAKE
...	...	...	...
31128	NFL Preview: Championship Match-Ups Prove Team...	The NFL is a league, so it should come as no...	REAL
31129	President Trump’s Father’s Day Proclamation: D...	President Donald Trump officially declared tod...	REAL
31130	Former Ambassador Andrew Young Calls for End t...	By Brandon Turbeville Anti-fluoridation activi...	FAKE
31131	Osama bin Laden’s older brother rents out luxu...	Osama bin Laden’s older brother rents out luxu...	FAKE
31132	WORLD WAR 3 ▲ Mr.President #004 ▲ xFrozenLPx	source Add To The Conversation Using Facebook ...	FAKE
31133	HUMA ABEDIN SWORE UNDER OATH SHE GAVE UP ‘ALL ...	Home › POLITICS	US NEWS › HUMA ABEDIN SWORE ...	FAKE

	Headline	Body	Label
31134	NaN	DYN's Statement on Last Week's Botnet Attack by...	FAKE
31135	NaN	Kinda reminds me of when Carter gave away the ...	FAKE
31137	Government Report: Islamists Building 'Paralle...	Aided by a politically correct culture of "tol...	REAL
31140	Editor of Austria's Largest Paper Charged with...	Breitbart October 26, 2016 \nAn editor of Aust...	FAKE
31141	This Is a Jobs Report That Democrats Can Boast...	There's not much to say about the July jobs nu...	REAL
31142	Christians in 2017 'Most Persecuted Group in t...	In many parts of the world, Christians gatheri...	REAL
31143	Florida Woman Charged in Death of Infant in 'C...	Early on Oct. 6, Erin was awakened by the so...	REAL
31144	Time is Running Out to Stop Kratom Ban – Need ...	By Brandon Turbeville When the DEA announced t...	FAKE
31145	The Fix Is In: NBC Affiliate Accidentally Post...	Home » Headlines » World News » The Fix Is In:...	FAKE
31146	Samsung, Kim Jong-un, Rex Tillerson: Your Morn...	Good morning. Here's what you need to know: •...	REAL
31147	Comment on World Heaves Sigh of Relief after T...	Finian Cunningham has written extensively on...	FAKE
31148	Ann Coulter: How to Provide Universal Health C...	The first sentence of Congress' Obamacare repe...	REAL
31149	Government Forces Advancing at Damascus-Aleppo...	#FROMTHEFRONT #MAPS 22.11.2016 - 1,361 views 5...	FAKE
31150	Sally Yates Won't Say If Trump Was Wiretapped ...	Former Deputy Attorney General Sally Yates dec...	REAL
31151	Maine's Gov. LePage Threatens To 'Investigate'...	Google Pinterest Digg Linkedin Reddit Stumbleu...	FAKE
31152	Sen. McConnell: The Supreme Court Vacancy Was ...	Senate Majority Leader Mitch McConnell (R, KY)...	REAL
31153	Nikki Haley Blasts U.N. Human Rights Office fo...	U. S Ambassador to the United Nations Nikki Ha...	REAL
31155	Jakarta Bombing Kills Three Police Officers, L...	Two suicide bombers attacked a bus station in ...	REAL
31156	Idiot Who Destroyed Trump Hollywood Star Gets ...	Share This \nAlthough the vandal who thought i...	FAKE
31157	Trump: Putin 'Very Smart' to Not Retaliate ove...	Donald Trump took to Twitter Friday to praise ...	REAL
31158	Rapper T.I.: Trump a 'Poster Child For White S...	Rapper T. I. unloaded on black celebrities who...	REAL
31159	N.F.L. Playoffs: Schedule, Matchups and Odds -...	When the Green Bay Packers lost to the Washing...	REAL
31160	Macy's Is Said to Receive Takeover Approach by...	The Macy's of today grew from the union of sev...	REAL
31162	What Keeps the F-35 Alive	David Swanson is an author, activist, journa...	FAKE

27865 rows × 3 columns

In [14]:

```
df['Label'].value_counts()
```

Out[14]:

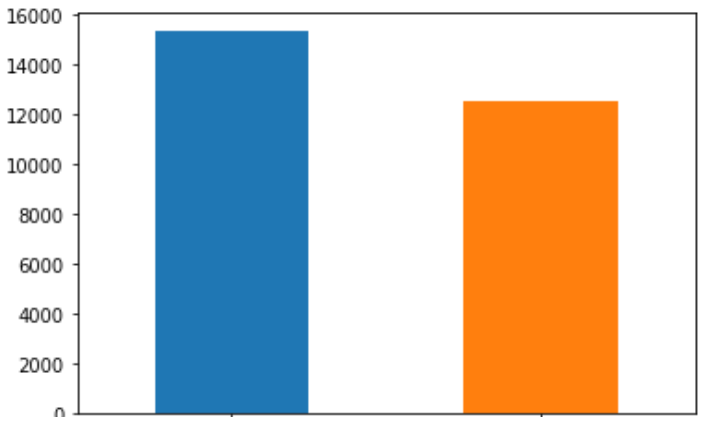
```
REAL    15343
FAKE    12522
Name: Label, dtype: int64
```

In [15]:

```
df['Label'].value_counts().plot(kind = 'bar')
```

Out[15]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7efe8a634ef0>
```



REAL - FAKE -

In [16]:

```
df['headline_length'] = [len(str(a)) for a in df['Headline']]
df['headline_length'].describe()
```

Out[16]:

```
count    27865.000000
mean        69.775381
std         24.885773
min          1.000000
25%         55.000000
50%         70.000000
75%         85.000000
max        653.000000
Name: headline_length, dtype: float64
```

In [17]:

```
df['body_length'] = [len(a) for a in df['Body']]
df['body_length'].describe()
```

Out[17]:

```
count     27865.000000
mean       4429.890903
std        4854.862554
min          10.000000
25%        1589.000000
50%        3348.000000
75%        6106.000000
max      142961.000000
Name: body_length, dtype: float64
```

In [18]:

```
df.describe()
```

Out[18]:

	headline_length	body_length
count	27865.000000	27865.000000
mean	69.775381	4429.890903
std	24.885773	4854.862554
min	1.000000	10.000000
25%	55.000000	1589.000000
50%	70.000000	3348.000000
75%	85.000000	6106.000000
max	653.000000	142961.000000

In [19]:

```
df["Text"] = df["Headline"].map(str) + df["Body"]
y = df.Label
y = y.astype('str')
X_train, X_test, Y_train, Y_test = train_test_split(df['Text'],y, test_size=0.33)
X_train
```

Out[19]:

```
4258     Donald Trump's GOP civil warPanama City, Flori...
24466    What You Should Watch: Amazon Pilots and 'Miss...
28155    Part 4 Of O'Keefe's Project Veritas Videos Has...
28266    Misophonia Sufferers: Scientists May Have Foun...
```