FAKE NEWS DETECTION

Project Title: FAKE NEWS DETECTION

Phase 3: Development Part 1

Problem Statement :

Fake News Classification with The Help Of Natural Language Processing Technique.Fake news detection is a hot topic in the field of natural language processing. We consume news through several mediums throughout the day in our daily routine, but sometimes it becomes difficult to decide which one is fake and which one is authentic. Our job is to create a model which predicts whether a given news is real or fake.

Required Libraries

```
import pandas as pd

import numpy as np

import re

import nltk

from nltk.corpus import stopwords

from nltk.stem import PorterStemmer, WordNetLemmatizer

from sklearn.model_selection import train_test_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.feature_extraction.text import TfidfVectorizer

from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
```

1. **Data Gathering :**

	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

2. **Data Analysis :**

```
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20800 entries, 0 to 20799
Data columns (total 5 columns):

 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   id      20800 non-null  int64
 1   title   20242 non-null  object
 2   author  18843 non-null  object
 3   text    20761 non-null  object
 4   label   20800 non-null  int64
dtypes: int64(2), object(3)
memory usage: 812.6+ KB

df['label'].value_counts()
1    10413
0    10387
Name: label, dtype: int64

df.shape
(20800, 5)

df.isna().sum()
id          0
title     558
author   1957
text       39
label       0
dtype: int64

df = df.dropna() #Handled Missing values by droping those rows

df.isna().sum()
id       0
title    0
```

```
author    0
text      0
label     0
dtype: int64

df.shape
(18285, 5)

df.reset_index(inplace=True)

df.head()
index  id  title  author  text  label
0  0  0  House Dem Aide: We Didn’t Even See Comey’s Let...
   Darrell Lucus  House Dem Aide: We Didn’t Even See Comey’s Let...  1
1  1  1  FLYNN: Hillary Clinton, Big Woman on Campus - ...  Daniel J.
Flynn  Ever get the feeling your life circles the rou...  0
2  2  2  Why the Truth Might Get You Fired  Consortiumnews.com
   Why the Truth Might Get You Fired October 29, ...  1
3  3  3  15 Civilians Killed In Single US Airstrike Hav...  Jessica
Purkiss  Videos 15 Civilians Killed In Single US Airstr...  1
4  4  4  Iranian woman jailed for fictional unpublished...  Howard
Portnoy  Print \r\nAn Iranian woman has been sentenced ...  1

df['title'][0]
'House Dem Aide: We Didn’t Even See Comey’s Letter Until Jason Chaffetz Tweeted
It'

df = df.drop(['id','text','author'],axis = 1)

df.head()
index  title  label
0  0  House Dem Aide: We Didn’t Even See Comey’s Let...  1
1  1  FLYNN: Hillary Clinton, Big Woman on Campus - ...  0
2  2  Why the Truth Might Get You Fired  1
3  3  15 Civilians Killed In Single US Airstrike Hav...  1
4  4  Iranian woman jailed for fictional unpublished...  1
```

3. Data Preprocessing :

1.Tokenization

```
sample_data = 'The quick brown fox jumps over the lazy dog'
```

```
sample_data = sample_data.split()
sample_data
['The', 'quick', 'brown', 'fox', 'jumps', 'over', 'the', 'lazy', 'dog']
```

2. Make Lowercase

```
sample_data = [data.lower() for data in sample_data]
sample_data
['the', 'quick', 'brown', 'fox', 'jumps', 'over', 'the', 'lazy', 'dog']
```

3. Remove Stopwords

```
stopwords = stopwords.words('english')
print(stopwords[0:10])
print(len(stopwords))
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're"]
179
sample_data = [data for data in sample_data if data not in stopwords]
print(sample_data)
len(sample_data)
['quick', 'brown', 'fox', 'jumps', 'lazy', 'dog']
6
```

4. Stemming

```
ps = PorterStemmer()
sample_data_stemming = [ps.stem(data) for data in sample_data]
print(sample_data_stemming)
['quick', 'brown', 'fox', 'jump', 'lazi', 'dog']
```

5. Lemmatization

```
lm = WordNetLemmatizer()
sample_data_lemma = [lm.lemmatize(data) for data in sample_data]
print(sample_data_lemma)
['quick', 'brown', 'fox', 'jump', 'lazy', 'dog']
```