Importing the Libraries

In [1]:

```
import numpy as np
import pandas as pd
import re #to hel[p then serarching the text in give datas
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

In [2]:

```
# This Python 3 environment comes with many helpful analytics libra
# It is defined by the kaggle/python Docker image: https://github.c
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_c

# Input data files are available in the read-only "../input/" direc
# For example, running this (by clicking run or pressing Shift+Ente

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/workin
# You can also write temporary files to /kaggle/temp/, but they won
```

```
/kaggle/input/fake-news/submit.csv
/kaggle/input/fake-news/train.csv
/kaggle/input/fake-news/test.csv
```

In [3]:

```
import nltk  # Import the Natural Language Toolkit library

nltk.download('stopwords')  # Download the stopwords dataset from N
```

```
[nltk_data] Downloading package stopwords to /usr/share/nltk_dat
a...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[3]:

```
True
```

In [4]:
```
# printing the stopwords in English
print(stopwords.words('english'))
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'yo
u', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yours
elf', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's",
'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', '
them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'w
hom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'ar
e', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'h
aving', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'bu
t', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'b
y', 'for', 'with', 'about', 'against', 'between', 'into', 'throug
h', 'during', 'before', 'after', 'above', 'below', 'to', 'from', '
up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', '
further', 'then', 'once', 'here', 'there', 'when', 'where', 'why',
'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'othe
r', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 's
o', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don',
"don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're',
've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn',
"didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't",
'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'm
ustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn',
"shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't",
'wouldn', "wouldn't"]
```

Data Pre-processing

In [5]:
```
# loading the dataset to a pandas DataFrame
news_dataset = pd.read_csv('/kaggle/input/fake-news/train.csv')
```

In [6]:
```
# Get the dimensions (number of rows and columns) of the news_datas
news_dataset.shape
```

Out[6]:
```
(20800, 5)
```

About the Dataset

id: unique id for a news article

title: the title of a news article

author: author of the news article

text: the text of the article; could be incomplete

label: a label that marks whether the news article is real or fake:

1: Fake news 0: real News

In [7]:

```
# print the first 5 rows of the dataframe
news_dataset.head()
```

Out[7]:

	id	title	author	text	label
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	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1

In [8]:

```
# counting the number of missing values in the dataset
news_dataset.isnull().sum()
```

Out[8]:

```
id           0
title      558
author    1957
text        39
label        0
dtype: int64
```

In [9]:

```
# replacing the null values with empty string
news_dataset = news_dataset.fillna('')
```

In [10]:

```
# merging the author name and news title
news_dataset['content'] = news_dataset['author']+' '+news_dataset['
```

In [11]:

```python
print(news_dataset['content'])
```

```
0        Darrell Lucus House Dem Aide: We Didn’t Even S...
1        Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo...
2        Consortiumnews.com Why the Truth Might Get You...
3        Jessica Purkiss 15 Civilians Killed In Single ...
4        Howard Portnoy Iranian woman jailed for fictio...
                               ...
20795    Jerome Hudson Rapper T.I.: Trump a ’Poster Chi...
20796    Benjamin Hoffman N.F.L. Playoffs: Schedule, Ma...
20797    Michael J. de la Merced and Rachel Abrams Macy...
20798    Alex Ansary NATO, Russia To Hold Parallel Exer...
20799              David Swanson What Keeps the F-35 Alive
Name: content, Length: 20800, dtype: object
```

In [12]:

```python
# separating the data & label
X = news_dataset.drop(columns='label', axis=1)
Y = news_dataset['label']
```

In [13]:

```python
print(X)
print(Y)
```

```
          id                                             title  \
0          0  House Dem Aide: We Didn’t Even See Comey’s Let...
1          1  FLYNN: Hillary Clinton, Big Woman on Campus - ...
2          2                 Why the Truth Might Get You Fired
3          3  15 Civilians Killed In Single US Airstrike Hav...
4          4  Iranian woman jailed for fictional unpublished...
...      ...                                               ...
20795  20795  Rapper T.I.: Trump a ’Poster Child For White S...
20796  20796  N.F.L. Playoffs: Schedule, Matchups and Odds -...
20797  20797  Macy’s Is Said to Receive Takeover Approach by...
20798  20798  NATO, Russia To Hold Parallel Exercises In Bal...
20799  20799                         What Keeps the F-35 Alive

                                          author  \
0                                  Darrell Lucus
1                                Daniel J. Flynn
2                             Consortiumnews.com
3                                Jessica Purkiss
4                                 Howard Portnoy
...                                          ...
20795                              Jerome Hudson
20796                           Benjamin Hoffman
20797  Michael J. de la Merced and Rachel Abrams
20798                                Alex Ansary
20799                              David Swanson

                                                    text  \
0      House Dem Aide: We Didn’t Even See Comey’s Let...
1      Ever get the feeling your life circles the rou...
2      Why the Truth Might Get You Fired October 29, ...
```

```
3      Videos 15 Civilians Killed In Single US Airstr...
4      Print \nAn Iranian woman has been sentenced to...
...                                                  ...
20795  Rapper T. I. unloaded on black celebrities who...
20796  When the Green Bay Packers lost to the Washing...
20797  The Macy’s of today grew from the union of sev...
20798  NATO, Russia To Hold Parallel Exercises In Bal...
20799    David Swanson is an author, activist, journa...

                                                 content
0      Darrell Lucus House Dem Aide: We Didn’t Even S...
1      Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo...
2      Consortiumnews.com Why the Truth Might Get You...
3      Jessica Purkiss 15 Civilians Killed In Single ...
4      Howard Portnoy Iranian woman jailed for fictio...
...                                                  ...
20795  Jerome Hudson Rapper T.I.: Trump a ’Poster Chi...
20796  Benjamin Hoffman N.F.L. Playoffs: Schedule, Ma...
20797  Michael J. de la Merced and Rachel Abrams Macy...
20798  Alex Ansary NATO, Russia To Hold Parallel Exer...
20799            David Swanson What Keeps the F-35 Alive

[20800 rows x 5 columns]
0        1
1        0
2        1
3        1
4        1
        ..
20795    0
20796    0
20797    0
20798    1
20799    1
Name: label, Length: 20800, dtype: int64
```

Stemming

Stemming is the process of reducing a word to its Root word

example: actor, actress, acting --> act remove prefix,suffix

In [14]:

```
# Create an instance of the PorterStemmer class
port_stem = PorterStemmer()
```

In [15]:
```
def stemming(content):
    # Remove all non-alphabetic characters and replace them with sp
    stemmed_content = re.sub('[^a-zA-Z]', ' ', content)

    # Convert the text to lowercase
    stemmed_content = stemmed_content.lower()

    # Split the text into individual words
    stemmed_content = stemmed_content.split()

    # Stem each word and remove stopwords
    stemmed_content = [port_stem.stem(word) for word in stemmed_con

    # Join the words back into a single string
    stemmed_content = ' '.join(stemmed_content)

    return stemmed_content  # Return the processed text
```

In [16]:
```
# Apply the stemming function to the 'content' column of the news_d
news_dataset['content'] = news_dataset['content'].apply(stemming)
```

In [17]:
```
print(news_dataset['content'])
```

```
0        darrel lucu hous dem aid even see comey letter...
1        daniel j flynn flynn hillari clinton big woman...
2                    consortiumnew com truth might get fire
3        jessica purkiss civilian kill singl us airstri...
4        howard portnoy iranian woman jail fiction unpu...
                               ...
20795    jerom hudson rapper trump poster child white s...
20796    benjamin hoffman n f l playoff schedul matchup...
20797    michael j de la merc rachel abram maci said re...
20798    alex ansari nato russia hold parallel exercis ...
20799                         david swanson keep f aliv
Name: content, Length: 20800, dtype: object
```

In [18]:
```
#separating the data and label
X = news_dataset['content'].values
Y = news_dataset['label'].values
```

In [19]:
```
print(X)
```

```
['darrel lucu hous dem aid even see comey letter jason chaffetz tw
eet'
 'daniel j flynn flynn hillari clinton big woman campu breitbart'
 'consortiumnew com truth might get fire' ...
 'michael j de la merc rachel abram maci said receiv takeov approa
ch hudson bay new york time'
 'alex ansari nato russia hold parallel exercis balkan'
 'david swanson keep f aliv']
```

In [20]:
```
print(Y)
```

```
[1 0 1 ... 0 1 1]
```

In [21]:
```
Y.shape
```

Out[21]:
```
(20800,)
```

In [22]:
```
# converting the textual data to numerical data
vectorizer = TfidfVectorizer()
vectorizer.fit(X)

X = vectorizer.transform(X)
```

In [23]:
```
print(X)
```

```
  (0, 15686)	0.28485063562728646
  (0, 13473)	0.2565896679337957
  (0, 8909)	0.3635963806326075
  (0, 8630)	0.29212514087043684
  (0, 7692)	0.24785219520671603
  (0, 7005)	0.21874169089359144
  (0, 4973)	0.233316966909351
  (0, 3792)	0.2705332480845492
  (0, 3600)	0.3598939188262559
  (0, 2959)	0.2468450128533713
  (0, 2483)	0.3676519686797209
  (0, 267)	0.27010124977708766
  (1, 16799)	0.30071745655510157
  (1, 6816)	0.1904660198296849
  (1, 5503)	0.7143299355715573
  (1, 3568)	0.26373768806048464
  (1, 2813)	0.19094574062359204
  (1, 2223)	0.3827320386859759
  (1, 1894)	0.15521974226349364
  (1, 1497)	0.2939891562094648
  (2, 15611)	0.41544962664721613
  (2, 9620)	0.49351492943649944
  (2, 5968)	0.3474613386728292
  (2, 5389)	0.3866530551182615
  (2, 3103)	0.46097489583229645
  :     :
  (20797, 13122)        0.2482526352197606
  (20797, 12344)        0.27263457663336677
  (20797, 12138)        0.24778257724396507
  (20797, 10306)        0.08038079000566466
  (20797, 9588) 0.174553480255222
  (20797, 9518) 0.2954204003420313
  (20797, 8988) 0.36160868928090795
  (20797, 8364) 0.22322585870464118
  (20797, 7042) 0.21799048897828688
  (20797, 3643) 0.21155500613623743
  (20797, 1287) 0.33538056804139865
  (20797, 699)  0.30685846079762347
  (20797, 43)   0.29710241860700626
  (20798, 13046)        0.22363267488270608
  (20798, 11052)        0.4460515589182236
  (20798, 10177)        0.3192496370187028
  (20798, 6889) 0.32496285694299426
  (20798, 5032) 0.4083701450239529
  (20798, 1125) 0.4460515589182236
  (20798, 588)  0.3112141524638974
  (20798, 350)  0.28446937819072576
  (20799, 14852)        0.5677577267055112
  (20799, 8036) 0.45983893273780013
  (20799, 3623) 0.37927626273066584
  (20799, 377)  0.5677577267055112
```

In [24]:

```
# Split the data into training and testing sets
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size
```

Training the Model: Logistic Regression

In [25]:

```
# Create an instance of the LogisticRegression model
model = LogisticRegression()
```

In [26]:

```
# Train the Logistic Regression model using the training data
model.fit(X_train, Y_train)
```

Out[26]:

```
LogisticRegression()
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

Evaluation

Accuracy score

In [27]:

```
# accuracy score on the training data
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train
```

In [28]:

```
print('Accuracy score of the training data : ', training_data_accur
```

```
Accuracy score of the training data :  0.9865985576923076
```

In [29]:

```
# accuracy score on the test data
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
```

In [30]:

```
print('Accuracy score of the test data : ', test_data_accuracy)
```

```
Accuracy score of the test data :  0.9790865384615385
```

Making a Predictive System

In [31]:

```
X_new = X_test[4]

prediction = model.predict(X_new)
print(prediction)

if (prediction[0]==0):
  print('The news is Real')
else:
  print('The news is Fake')
```

```
[0]
The news is Real
```

check ouput is corrct/wrong to compare actual output)

In [32]:

```
print(Y_test[4])
```

```
0
```

In [33]:

```
import pickle
```

In [34]:

```
# Save the trained model to a file
with open("fake_news_prediction_model.pkl", "wb") as file:
    pickle.dump(model, file)
```

In [ ]: