Fake News Detection

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [4]:
```
data = pd.read_csv("fake_news.csv")
data.head()
```

Out[4]:

	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4	4	Iranian woman jailed for fictional unpublished...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 [5]:
```
data.shape
```

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

In [6]:
```
data.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
```

In [7]:
```
data.isna().sum()
```

Out[7]:
```
id          0
title     558
author   1957
text       39
label       0
dtype: int64
```

In [8]:
```
data = data.drop(['id'], axis=1)
```

In [9]:
```
data = data.fillna('')
```

In [10]:
```
data['content'] = data['author']+' '+data['title']+' '+data['text']
```

In [11]:
```
data = data.drop(['title','author', 'text'], axis=1)
```

In [12]:
```
data.head()
```

Out[12]:

	label	content
0	1	Darrell Lucus House Dem Aide: We Didn’t Even S...
1	0	Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo...
2	1	Consortiumnews.com Why the Truth Might Get You...
3	1	Jessica Purkiss 15 Civilians Killed In Single ...
4	1	Howard Portnoy Iranian woman jailed for fictio...

In [13]:
```
data['content'] = data['content'].apply(lambda x: " ".join(x.lower() for x in x.split())
```

In [14]:
```
data['content'] = data['content'].str.replace('[^\w\s]','')
```

```
C:\Users\Pooja Reddy\AppData\Local\Temp\ipykernel_17260\3643324700.py:1: FutureWarning:
The default value of regex will change from True to False in a future version.
  data['content'] = data['content'].str.replace('[^\w\s]','')
```

In [15]:
```
import nltk
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to C:\Users\Pooja
[nltk_data]     Reddy\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[15]:
```
True
```

In [37]:
```
import nltk
nltk.download('wordnet')
```

```
[nltk_data] Downloading package wordnet to C:\Users\Pooja
[nltk_data]     Reddy\AppData\Roaming\nltk_data...
```

Out[37]:
```
True
```

In [38]:
```
from nltk.corpus import stopwords
stop = stopwords.words('english')
data['content'] = data['content'].apply(lambda x: " ".join(x for x in x.split() if x not
```

In [*]:
```
from nltk.stem import WordNetLemmatizer
from textblob import Word
data['content'] = data['content'].apply(lambda x: " ".join([Word(word).lemmatize() for w
data['content'].head()
```

In [*]:
```
X = data[['content']]
y = data['label']
```

In [*]:
```
from sklearn.model_selection import train_test_split
```

In [*]:
```
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3, random_state=45, strat
```

In [*]:
```
print (X_train.shape)
print (y_train.shape)
print (X_test.shape)
print (y_test.shape)
```

In [*]:
```
from sklearn.feature_extraction.text import TfidVectorizer
```

In [*]:
```
from sklearn.feature_extraction.text import TfidfVectorizer

tfidf_vect = TfidfVectorizer(analyzer='word', token_pattern=r'\w{1,}', max_features=5000
tfidf_vect.fit(data['content'])

xtrain_tfidf = tfidf_vect.transform(X_train['content'])
xtest_tfidf = tfidf_vect.transform(X_test['content'])
```

In [ ]:
```

```