In [ ]:
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.pipeline import make_pipeline
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import classification_report, accuracy_score

import numpy as np
import math
import pandas as pd
import re
import nltk
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize
import warnings
import seaborn as sns
import matplotlib.pyplot as plt
warnings.filterwarnings("ignore")
```

In [ ]:
```
df = pd.read_csv("WELFake_Dataset.csv")
```

In [ ]:
```
df = df.fillna('')
```

In [ ]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 72134 entries, 0 to 72133
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Unnamed: 0  72134 non-null  int64 
 1   title       72134 non-null  object
 2   text        72134 non-null  object
 3   label       72134 non-null  int64 
dtypes: int64(2), object(2)
memory usage: 2.2+ MB
```

In [ ]:
```
df.drop(columns=['Unnamed: 0'], inplace=True)
```

In [ ]:
```
df
```

Out[ ]:

	title	text	label
0	LAW ENFORCEMENT ON HIGH ALERT Following Threat...	No comment is expected from Barack Obama Membe...	1
1		Did they post their votes for Hillary already?	1
2	UNBELIEVABLE! OBAMA'S ATTORNEY GENERAL SAYS MO...	Now, most of the demonstrators gathered last ...	1
3	Bobby Jindal, raised Hindu, uses story of Chri...	A dozen politically active pastors came here f...	0
4	SATAN 2: Russia unvelis an image of its terrif...	The RS-28 Sarmat missile, dubbed Satan 2, will...	1
...	...	...	...
72129	Russians steal research on Trump in hack of U....	WASHINGTON (Reuters) - Hackers believed to be ...	0
72130	WATCH: Giuliani Demands That Democrats Apolog...	You know, because in fantasyland Republicans n...	1
72131	Migrants Refuse To Leave Train At Refugee Camp...	Migrants Refuse To Leave Train At Refugee Camp...	0
72132	Trump tussle gives unpopular Mexican leader mu...	MEXICO CITY (Reuters) - Donald Trump's combati...	0
72133	Goldman Sachs Endorses Hillary Clinton For Pre...	Goldman Sachs Endorses Hillary Clinton For Pre...	1

72134 rows × 3 columns

In [ ]:
```
df.isnull().value_counts()
```

Out[ ]:
```
title  text   label
False  False  False    72134
Name: count, dtype: int64
```

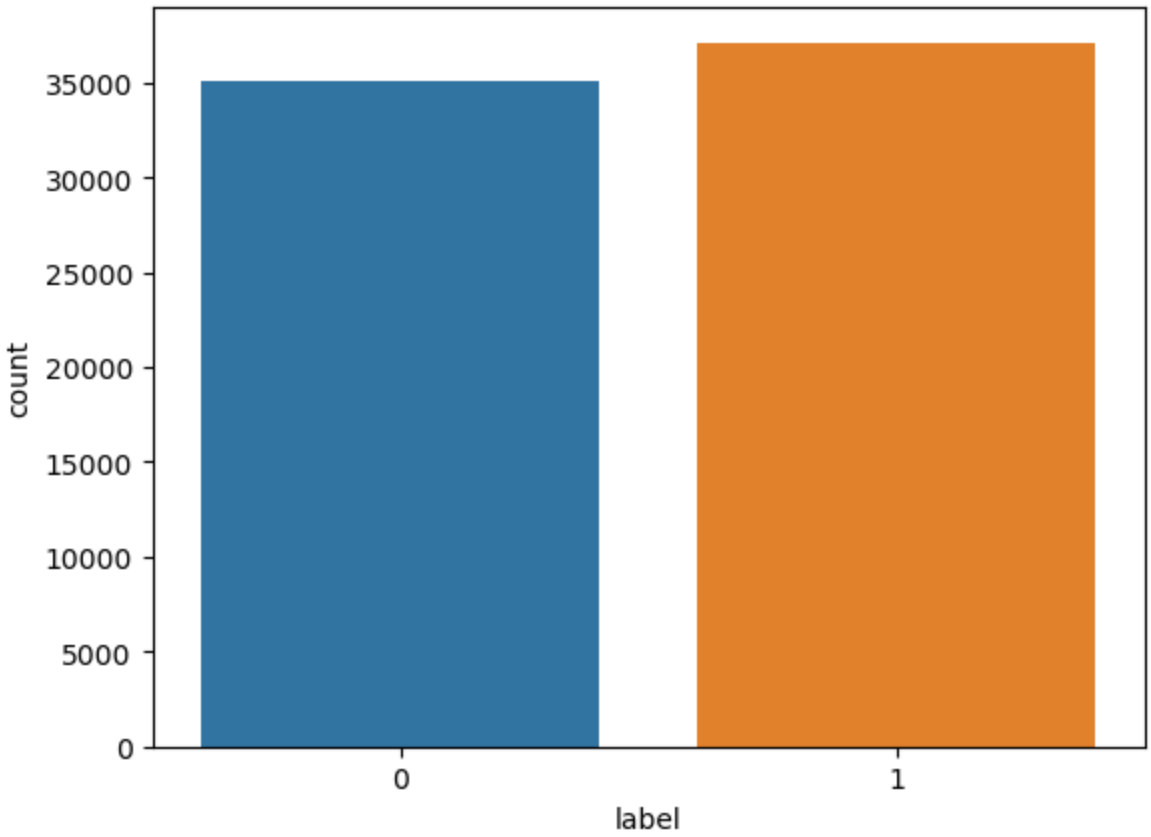
In [ ]:
```
df.isnull().sum()
```

Out[ ]:
```
title    0
text     0
label    0
dtype: int64
```

In [ ]:
```
import seaborn as sns

sns.countplot(data=df, x='label')
```

Out[ ]:
```
<Axes: xlabel='label', ylabel='count'>
```



In [ ]:
```
df['label'].value_counts()
```

Out[ ]:
```
label
1    37106
0    35028
Name: count, dtype: int64
```

In [ ]:
```
for col in df.columns:
    print(col, ':', df[col].nunique())
```

```
title : 62348
text : 62719
label : 2
```

In [ ]:
```
any_nan = df.isnull().any().any()

# Display the result
print(f"Are there any NaN values in the DataFrame? {any_nan}")
df
```

```
Are there any NaN values in the DataFrame? False
```

Out[ ]:

	title	text	label
0	LAW ENFORCEMENT ON HIGH ALERT Following Threat...	No comment is expected from Barack Obama Membe...	1
1		Did they post their votes for Hillary already?	1
2	UNBELIEVABLE! OBAMA'S ATTORNEY GENERAL SAYS MO...	Now, most of the demonstrators gathered last ...	1
3	Bobby Jindal, raised Hindu, uses story of Chri...	A dozen politically active pastors came here f...	0
4	SATAN 2: Russia unvelis an image of its terrif...	The RS-28 Sarmat missile, dubbed Satan 2, will...	1
...	...	...	...
72129	Russians steal research on Trump in hack of U....	WASHINGTON (Reuters) - Hackers believed to be ...	0
72130	WATCH: Giuliani Demands That Democrats Apolog...	You know, because in fantasyland Republicans n...	1
72131	Migrants Refuse To Leave Train At Refugee Camp...	Migrants Refuse To Leave Train At Refugee Camp...	0
72132	Trump tussle gives unpopular Mexican leader mu...	MEXICO CITY (Reuters) - Donald Trump's combati...	0
72133	Goldman Sachs Endorses Hillary Clinton For Pre...	Goldman Sachs Endorses Hillary Clinton For Pre...	1

72134 rows × 3 columns

In [ ]:
```
# Combining title and text
df['content'] = df['title'] + ' ' + df['text']

# Drop rows with missing values
df.dropna(inplace=True)
```

Upsampling the minority class

It is known that Naive bayes is not robust to class imbalance. It could be seen above that the data is little imbalanced. Therefore, class balancing can be done before giving it to the Naive Bayes model for prediction.

Feel free to use 'resample' library from sklearn.

In [ ]:
```
from sklearn.utils import resample

df_majority = df[df['label']==1]
df_minority = df[df['label']==0]
```

```
negative_upsample = resample(df_minority, replace = True,
                        n_samples = df_majority.shape[0],
                        random_state = 101)
df_upsampled = pd.concat([df_majority, negative_upsample])  # concat two dat
df_upsampled = df_upsampled.sample(frac = 1)
```

In [ ]:
```
df_upsampled[df_upsampled['label']==0].shape
```

Out[ ]:
```
(37106, 4)
```

In [ ]:
```
## In this cell, we are going to be dividing the data into train and test pc
## Ensure that you store the upsampled data in a variable called 'df_upsampl
## so that the below operations are performed successfully


## Considering 10000 positive and 10000 negative data points
negative_data_points_train = df_upsampled[df_upsampled['label']==0].iloc[:29
positive_data_points_train = df_upsampled[df_upsampled['label']==1].iloc[:29

## Considering the remaining data points for test
negative_data_points_test = df_upsampled[df_upsampled['label']==0].iloc[2900
positive_data_points_test = df_upsampled[df_upsampled['label']==1].iloc[2900
## Concatenate the training positive and negative contents
X_train = pd.concat([positive_data_points_train['content'], negative_data_pc
## Concatenating the training positive and negative outputs
y_train = pd.concat([positive_data_points_train['label'], negative_data_poir

## Concatenating the test positive and negative contents
X_test = pd.concat([positive_data_points_test['content'], negative_data_poir
## Concatenating the test positive and negative outputs
y_test = pd.concat([positive_data_points_test['label'], negative_data_points
```

In [ ]:
```
y_train.value_counts()
```

Out[ ]:
```
label
1    29000
0    29000
Name: count, dtype: int64
```

In [ ]:
```
y_test.value_counts()
```

Out[ ]:
```
label
1    8106
0    8106
Name: count, dtype: int64
```

Pre-process the reviews:

We know that a review contains links, punctuation, stopwords and many other words that don't give a lot of meaning for the Naive Bayes model for prediction.

In the cell below, we implement text-preprocessing and remove links, punctuations and stopwords. It is also important to lowercase the letters so that 'Admire' and 'admire' are

not treated as different words.

In addition to this, perform stemming operation so that similar words are reduced.

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

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/deepthikondragunta/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

Out[ ]: True

In [ ]:
```
# TASK CELL
import re
import string

def remove_stopwords(text):
    temp=[]
    for word in text.split():
        if word in stopwords.words('english'):
            temp.append('')
        else:
            temp.append(word)
    x=temp[:]
    temp.clear()
    return " ".join(x)

def clean_review(review):
    '''
    Input:
        review: a string containing a review.
    Output:
        review_cleaned: a processed review.

    '''
    review_cleaned=review.lower() #converting the reviews to lowercase
    review_cleaned=re.sub(r'<.*?>','',review_cleaned)  #the html tags are re
    review_cleaned=re.sub(r'http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\\(\\)
    #removing punctuations form the text
    exclude=string.punctuation
    review_cleaned=review_cleaned.translate(str.maketrans('','',exclude))
    review_cleaned=remove_stopwords(review_cleaned) #calling an external fur
    ps=PorterStemmer() #initializing the porter stemmer object
    review_cleaned=" ".join([ps.stem(word) for word in review_cleaned.split(
    return review_cleaned
```

In [ ]:
```
X_train.iloc[0]
```

Out[ ]:
```
' Watch Out, Clarence Thomas: Petition Asks President Obama To Nominate Ani
ta Hill For SCOTUS Conservatives would be so pissed if this actually happen
ed.In what would perhaps be the most entertaining Supreme Court nominating
process in American history, a petition is circulating asking President Oba
ma to nominate Anita Hill to replace recently deceased Justice Antonin Scal
ia on the bench.Hill is most remembered for her courageous testimony agains
t current Supreme Court Justice Clarence Thomas during his confirmation hea
rings in 1991. Hill testified that Thomas made unwanted sexual advances tow
ard her during his stint as supervisor at the Department of Education. Desp
ite her passing a lie detector test while he refused to take one, the Senat
e still confirmed Thomas 52-48 in the narrowest margin since the 1800s afte
r other women were denied the chance to testify in support of Hill.Thomas a
nd his conservative supporters, of course, demonized Hill, accusing her of
being used by white liberals to cut down an  uppity black  with a  high-tec
h lynching. But while Hill may seem to be a controversial choice to fill Sc
alia s seat on the high court, it s not out of the realm of possibility, no
r does she lack the qualifications.Hill is an experienced attorney who also
serves as University Professor of Social Policy, Law, and Women s Studies a
t Brandeis University. The 59-year-old attended Oklahoma State University a
nd Yale Law School. She is one of the most prominent experts in her field a
nd certainly possesses the legal and academic chops to serve on the Supreme
Court.Furthermore, there has never been an African-American woman on the Co
urt, which makes this an opportunity to make history with a much needed cha
nge. What better way to replace a racist misogynist like Scalia than with a
n educated black woman who specializes in social policy?Not only that, just
imagine how uncomfortable her nomination would make Clarence Thomas feel. H
e d probably be sweating bullets while watching and hoping the nomination p
rocess eliminates her as the nominee. And it would be incredibly hard for c
onservatives to grill her without reminding the American public of how big
of creep Thomas is. And if Republicans are too hard on her, they can be the
ones accused of a  high-tech lynching  of an  uppity black  woman as they l
et their sexism and racism fly during hearings that would would likely be n
ationally televised and strewn across social media.And even if Hill fails t
o be confirmed, it would make Republicans look like the terrible lawmakers
and human beings that they are, all while embarrassing the hell out of Thom
as, who may even end up feeling too exposed to remain on the Court. And if
she does get confirmed, he might resign anyway or at the very least be forc
ed to watch as the woman he harassed and humiliated over 20 years ago puts
on the same black robe to help the American people in a way he has refused
to do throughout his own tenure. She could end up being the social justice
crusader women and minorities have hoped for and become more revered than S
calia and Thomas could ever hope to be.As the petition says,  Now THAT S Ju
stice! Featured image from Wikimedia'
```

In [ ]:
```
custom_review = X_train.iloc[0]

# print cleaned review
print(clean_review(custom_review))
```

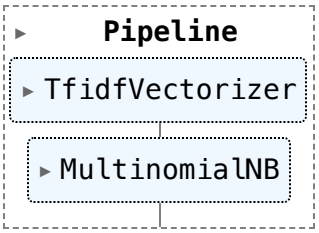
```
watch clarenc thoma petit ask presid obama nomin anita hill scotu conserv wo
uld piss actual happenedin would perhap entertain suprem court nomin process
american histori petit circul ask presid obama nomin anita hill replac recen
t deceas justic antonin scalia benchhil rememb courag testimoni current supr
em court justic clarenc thoma confirm hear 1991 hill testifi thoma made unwa
nt sexual advanc toward stint supervisor depart educ despit pass lie detecto
r test refus take one senat still confirm thoma 5248 narrowest margin sinc 1
800 women deni chanc testifi support hillthoma conserv support cours demon h
ill accus use white liber cut uppiti black hightech lynch hill may seem cont
roversi choic fill scalia seat high court realm possibl lack qualificationsh
il experienc attorney also serv univers professor social polici law women st
udi brandei univers 59yearold attend oklahoma state univers yale law school
one promin expert field certainli possess legal academ chop serv suprem cour
tfurthermor never africanamerican woman court make opportun make histori muc
h need chang better way replac racist misogynist like scalia educ black woma
n special social policynot imagin uncomfort nomin would make clarenc thoma f
eel probabl sweat bullet watch hope nomin process elimin nomine would incred
hard conserv grill without remind american public big creep thoma republican
hard one accus hightech lynch uppiti black woman let sexism racism fli hear
would would like nation televis strewn across social mediaand even hill fail
confirm would make republican look like terribl lawmak human be embarrass he
ll thoma may even end feel expos remain court get confirm might resign anywa
y least forc watch woman harass humili 20 year ago put black robe help ameri
can peopl way refus throughout tenur could end social justic crusad women mi
nor hope becom rever scalia thoma could ever hope bea petit say justic featu
r imag wikimedia
```

In [ ]:
```
# Apply clean_review function to the training data
X_train= X_train.apply(clean_review)
# Apply clean_review function to the test data
X_test = X_test.apply(clean_review)
```

In [ ]:
```
# X_train, X_test, y_train, y_test = train_test_split(df['content'], df['lab
```

In [ ]:
```
model = make_pipeline(TfidfVectorizer(), MultinomialNB())
model.fit(X_train, y_train)
```

Out[ ]:



In [ ]:
```
predicted = model.predict(X_test)
print(classification_report(y_test, predicted))
print("Accuracy:", round(accuracy_score(y_test, predicted)*100),'%')
```

```
                precision    recall  f1-score   support

           0       0.84      0.94      0.89      8106
           1       0.93      0.82      0.87      8106

    accuracy                           0.88     16212
   macro avg       0.89      0.88      0.88     16212
weighted avg       0.89      0.88      0.88     16212

Accuracy: 88 %
```

In [ ]:
```python
# def predict_fake_news(news):
#     prediction = model.predict([news])
#     return 'Fake' if prediction[0] == 0 else 'Real'
# # Example usage
# print(predict_fake_news("SATAN 2: Russia unvelis an image of its terrif.")

def predict_fake_news(news):
    # Clean the input news text
    cleaned_news = clean_review(news)

    # Make predictions using the model
    prediction = model.predict([cleaned_news])

    # Return the result as 'Fake' or 'Real'
    return 'Fake' if prediction[0] == 0 else 'Real'
```

In [ ]:
```python
# Example usage
news_text = "SATAN 2: Russia unveils an image of its terrif."
print(predict_fake_news(news_text))
```

```
Real
```

In [ ]:
```python
print(predict_fake_news("Ukraine is being invaded by russia."))
```

```
Real
```

In [ ]:
```python
print(predict_fake_news("India won the fifa worldcup."))
```

```
Fake
```

In [ ]:
```python

```