In [19]:
```
pip install pandas scikit-learn nltk
```

```
Requirement already satisfied: pandas in c:\anaconda\lib\site-packages (1.5.3)
Requirement already satisfied: scikit-learn in c:\anaconda\lib\site-packages (1.3.0)
Requirement already satisfied: nltk in c:\anaconda\lib\site-packages (3.8.1)
Requirement already satisfied: python-dateutil>=2.8.1 in c:\anaconda\lib\site-packages (from pandas) (2.8.
2)
Requirement already satisfied: pytz>=2020.1 in c:\anaconda\lib\site-packages (from pandas) (2022.7)
Requirement already satisfied: numpy>=1.21.0 in c:\anaconda\lib\site-packages (from pandas) (1.24.3)
Requirement already satisfied: scipy>=1.5.0 in c:\anaconda\lib\site-packages (from scikit-learn) (1.10.1)
Requirement already satisfied: joblib>=1.1.1 in c:\anaconda\lib\site-packages (from scikit-learn) (1.2.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\anaconda\lib\site-packages (from scikit-learn)
(2.2.0)
Requirement already satisfied: click in c:\anaconda\lib\site-packages (from nltk) (8.0.4)
Requirement already satisfied: regex>=2021.8.3 in c:\anaconda\lib\site-packages (from nltk) (2022.7.9)
Requirement already satisfied: tqdm in c:\anaconda\lib\site-packages (from nltk) (4.65.0)
Requirement already satisfied: six>=1.5 in c:\anaconda\lib\site-packages (from python-dateutil>=2.8.1->pand
as) (1.16.0)
Requirement already satisfied: colorama in c:\anaconda\lib\site-packages (from click->nltk) (0.4.6)
Note: you may need to restart the kernel to use updated packages.
```

In [20]:
```
import pandas as pd
import nltk
import string
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
```

In [21]:
```
nltk.download('punkt')
nltk.download('stopwords')
```

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

Out[21]: True

In [22]:
```
data = pd.read_csv("spam.csv", encoding="latin-1")
data
```

Out[22]:

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	ham	Go until jurong point, crazy.. Available only ...	NaN	NaN	NaN
1	ham	Ok lar... Joking wif u oni...	NaN	NaN	NaN
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	NaN	NaN	NaN
3	ham	U dun say so early hor... U c already then say...	NaN	NaN	NaN
4	ham	Nah I don't think he goes to usf, he lives aro...	NaN	NaN	NaN
...	...	...	...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...	NaN	NaN	NaN
5568	ham	Will Ì_ b going to esplanade fr home?	NaN	NaN	NaN
5569	ham	Pity, * was in mood for that. So...any other s...	NaN	NaN	NaN
5570	ham	The guy did some bitching but I acted like i'd...	NaN	NaN	NaN
5571	ham	Rofl. Its true to its name	NaN	NaN	NaN

In [23]:
```
data = data[['v1', 'v2']]
data.columns = ['label', 'text']
```

In [24]:
```
def process_text(text):
    tokens = word_tokenize(text)
    tokens = [word.lower() for word in tokens if word.isalnum()]
    tokens = [word for word in tokens if word not in stopwords.words('english')]
    return ' '.join(tokens)

data['text'] = data['text'].apply(process_text)
```

```
C:\Users\Muskan\AppData\Local\Temp\ipykernel_18700\1047370577.py:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html
#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#retu
rning-a-view-versus-a-copy)
  data['text'] = data['text'].apply(process_text)
```

In [25]:
```
X = data['text']
X
```

Out[25]:
```
0       go jurong point crazy available bugis n great ...
1                                 ok lar joking wif u oni
2       free entry 2 wkly comp win fa cup final tkts 2...
3                     u dun say early hor u c already say
4                   nah think goes usf lives around though
                              ...
5567    2nd time tried 2 contact u pound prize 2 claim...
5568                             b going esplanade fr home
5569                                 pity mood suggestions
5570    guy bitching acted like interested buying some...
5571                                       rofl true name
Name: text, Length: 5572, dtype: object
```

In [26]:
```
y = data['label']
y
```

Out[26]:
```
0        ham
1        ham
2       spam
3        ham
4        ham
        ...
5567    spam
5568     ham
5569     ham
5570     ham
5571     ham
Name: label, Length: 5572, dtype: object
```

In [27]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

In [28]:
```
tfidf_vectorizer = TfidfVectorizer()
X_train_tfidf = tfidf_vectorizer.fit_transform(X_train)
X_test_tfidf = tfidf_vectorizer.transform(X_test)
```

In [29]:
```
classifier = MultinomialNB()
classifier.fit(X_train_tfidf, y_train)
```

Out[29]:
```
MultinomialNB()
```

In [30]:
```
y_pred = classifier.predict(X_test_tfidf)
print(y_pred)
```

```
['ham' 'ham' 'ham' ... 'ham' 'ham' 'spam']
```

In [31]:
```
accuracy = accuracy_score(y_test, y_pred)
confusion = confusion_matrix(y_test, y_pred)
classification_rep = classification_report(y_test, y_pred)
```

In [32]:
```
print("Accuracy:", accuracy)
```

```
Accuracy: 0.967713004484305
```

In [33]:
```
print("Confusion Matrix:\n", confusion)
```

```
Confusion Matrix:
 [[965   0]
 [ 36 114]]
```

In [34]:
```
print("Classification Report:\n", classification_rep)
```

```
Classification Report:
              precision    recall  f1-score   support

         ham       0.96      1.00      0.98       965
        spam       1.00      0.76      0.86       150

    accuracy                           0.97      1115
   macro avg       0.98      0.88      0.92      1115
weighted avg       0.97      0.97      0.97      1115
```

In [ ]:

In [ ]: