Bharat Intern

Task - 1 (SMS Classifier)

Done by Cherukuri Krishna Lathvik

Importing the required libraries

In [9]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
```

Loading the Dataset

In [15]:
```
df = pd.read_csv('spam.csv', encoding='latin-1')
df.head()
```

Out[15]:

	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

In [16]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5572 entries, 0 to 5571
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   v1          5572 non-null   object
 1   v2          5572 non-null   object
 2   Unnamed: 2  50 non-null     object
 3   Unnamed: 3  12 non-null     object
 4   Unnamed: 4  6 non-null      object
dtypes: object(5)
memory usage: 217.8+ KB
```

In [17]:
```
df.describe
```

Out[17]:
```
<bound method NDFrame.describe of         v1                                                 v2 Unnamed: 2  \
0      ham  Go until jurong point, crazy.. Available only ...        NaN
1      ham                      Ok lar... Joking wif u oni...        NaN
2     spam  Free entry in 2 a wkly comp to win FA Cup fina...        NaN
3      ham  U dun say so early hor... U c already then say...        NaN
4      ham  Nah I don't think he goes to usf, he lives aro...        NaN
...    ...                                                ...        ...
5567  spam  This is the 2nd time we have tried 2 contact u...        NaN
5568   ham              Will Ì_ b going to esplanade fr home?        NaN
5569   ham  Pity, * was in mood for that. So...any other s...        NaN
5570   ham  The guy did some bitching but I acted like i'd...        NaN
5571   ham                        Rofl. Its true to its name        NaN

     Unnamed: 3 Unnamed: 4
0           NaN        NaN
1           NaN        NaN
2           NaN        NaN
3           NaN        NaN
4           NaN        NaN
...         ...        ...
5567        NaN        NaN
5568        NaN        NaN
5569        NaN        NaN
5570        NaN        NaN
5571        NaN        NaN

[5572 rows x 5 columns]>
```

In [25]:
```
df.shape
```

Out[25]:
```
(5572, 5)
```

Checking the Null values

In [18]:
```
df.isnull().sum()
```

Out[18]:
```
v1              0
v2              0
Unnamed: 2    5522
Unnamed: 3    5560
Unnamed: 4    5566
dtype: int64
```

In [20]:
```
df['v1'].value_counts()
```

Out[20]:
```
ham     4825
spam     747
Name: v1, dtype: int64
```

In [21]:
```
df['v2'].value_counts()
```

Out[21]:
```
Sorry, I'll call later
30
I cant pick the phone right now. Pls send a message
12
Ok...
10
7 wonders in My WORLD 7th You 6th Ur style 5th Ur smile 4th Ur Personality 3rd Ur
Nature 2nd Ur SMS and 1st \Ur Lovely Friendship\"... good morning dear"
4
Say this slowly.? GOD,I LOVE YOU &amp;amp; I NEED YOU,CLEAN MY HEART WITH YOUR BLOOD.S
end this to Ten special people &amp;amp; u c miracle tomorrow, do it,pls,pls do it...
4

..
I gotta collect da car at 6 lei.
1
No. On the way home. So if not for the long dry spell the season would have been o
ver
1
Urgent! Please call 09061743811 from landline. Your ABTA complimentary 4* Tenerife
Holiday or å£5000 cash await collection SAE T&Cs Box 326 CW25WX 150ppm
1
Dear 0776xxxxxxx U've been invited to XCHAT. This is our final attempt to contact
u! Txt CHAT to 86688 150p/MsgrcvdHG/Suite342/2Lands/Row/W1J6HL LDN 18yrs
1
Rofl. Its true to its name
1
Name: v2, Length: 5169, dtype: int64
```

Preprocessing the data

In [22]:
```
df['v1'] = df['v1'].map({'ham': 0, 'spam': 1}) #Mapping labels to 0 as non-spam and
```

Splitting the Data into Training and Testing Sets

In [23]:
```
X_train, X_test, y_train, y_test = train_test_split(df['v2'], df['v1'], test_size=0
```

In [24]:
```
X_train.shape
```

Out[24]:
```
(4457,)
```

Converting the text data

In [27]:
```
vectorizer = CountVectorizer()
X_train_vectorized = vectorizer.fit_transform(X_train)
X_test_vectorized = vectorizer.transform(X_test)
```

In [28]:
```
df.head()
```

Out[28]:

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	0	Go until jurong point, crazy.. Available only ...	NaN	NaN	NaN
1	0	Ok lar... Joking wif u oni...	NaN	NaN	NaN
2	1	Free entry in 2 a wkly comp to win FA Cup fina...	NaN	NaN	NaN
3	0	U dun say so early hor... U c already then say...	NaN	NaN	NaN
4	0	Nah I don't think he goes to usf, he lives aro...	NaN	NaN	NaN

Now training the model

In [30]:
```
classifier = MultinomialNB()
classifier.fit(X_train_vectorized, y_train)
```

Out[30]:
```
MultinomialNB()
```

Making predictions on the testing set

In [31]:
```
y_pred = classifier.predict(X_test_vectorized)
```

Evaluating the model

In [32]:
```
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)
classification_rep = classification_report(y_test, y_pred)

print(f"Accuracy: {accuracy}")
print("\nConfusion Matrix:")
print(conf_matrix)
print("\nClassification Report:")
print(classification_rep)
```

```
Accuracy: 0.9838565022421525

Confusion Matrix:
[[963   2]
 [ 16 134]]

Classification Report:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99       965
           1       0.99      0.89      0.94       150

    accuracy                           0.98      1115
   macro avg       0.98      0.95      0.96      1115
weighted avg       0.98      0.98      0.98      1115
```

In [ ]: