In [4]:

```
!pip install pandas scikit-learn nltk
```

```
Requirement already satisfied: pandas in c:\users\nalla\anaconda3\lib\site-packages (2.0.3)
Requirement already satisfied: scikit-learn in c:\users\nalla\anaconda3\lib\site-packages (1.3.0)
Requirement already satisfied: nltk in c:\users\nalla\anaconda3\lib\site-packages (3.8.1)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\nalla\anaconda3\lib\site-packages (from panda
s) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\nalla\anaconda3\lib\site-packages (from pandas) (2023.3
.post1)
Requirement already satisfied: tzdata>=2022.1 in c:\users\nalla\anaconda3\lib\site-packages (from pandas) (2023
.3)
Requirement already satisfied: numpy>=1.21.0 in c:\users\nalla\anaconda3\lib\site-packages (from pandas) (1.24.
3)
Requirement already satisfied: scipy>=1.5.0 in c:\users\nalla\anaconda3\lib\site-packages (from scikit-learn) (
1.11.1)
Requirement already satisfied: joblib>=1.1.1 in c:\users\nalla\anaconda3\lib\site-packages (from scikit-learn)
(1.2.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\nalla\anaconda3\lib\site-packages (from scikit-
learn) (2.2.0)
Requirement already satisfied: click in c:\users\nalla\anaconda3\lib\site-packages (from nltk) (8.0.4)
Requirement already satisfied: regex>=2021.8.3 in c:\users\nalla\anaconda3\lib\site-packages (from nltk) (2022.
7.9)
Requirement already satisfied: tqdm in c:\users\nalla\anaconda3\lib\site-packages (from nltk) (4.65.0)
Requirement already satisfied: six>=1.5 in c:\users\nalla\anaconda3\lib\site-packages (from python-dateutil>=2.
8.2->pandas) (1.16.0)
Requirement already satisfied: colorama in c:\users\nalla\anaconda3\lib\site-packages (from click->nltk) (0.4.6
)
```

In [7]:

```
import pandas as pd
data=pd.read_csv(r"C:\Users\nalla\Desktop\projects\spam.csv",encoding="latin1")
data=data[['v1','v2']]
data.columns=['label','message']
```

In [16]:

```
import re
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from sklearn.feature_extraction.text import TfidfVectorizer
import nltk
nltk.download('punkt')
nltk.download('stopwords')
stop_words=set(stopwords.words('english'))
def pre_process_text(text):
    text=re.sub('[^a-zA-Z]',' ',text)
    text=text.lower()
    words=word_tokenize(text)
    words=[word for word in words if word not in stop_words]
    return ' '.join(words)
data['message']=data['message'].apply(pre_process_text)
```

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

In [34]:

```
df.head(10)
```

Out[34]:

	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
5	spam	FreeMsg Hey there darling it's been 3 week's n...	NaN	NaN	NaN
6	ham	Even my brother is not like to speak with me. ...	NaN	NaN	NaN
7	ham	As per your request 'Melle Melle (Oru Minnamin...	NaN	NaN	NaN
8	spam	WINNER!! As a valued network customer you have...	NaN	NaN	NaN
9	spam	Had your mobile 11 months or more? U R entitle...	NaN	NaN	NaN

In [35]:

```
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 [9]:

```
df.describe()
```

Out[9]:

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
count	5572	5572	50	12	6
unique	2	5169	43	10	5
top	ham	Sorry, I'll call later	bt not his girlfrnd... G o o d n i g h t . . .@"	MK17 92H. 450Ppw 16"	GNT:-)"
freq	4825	30	3	2	2

In [10]:

```
df.isnull().sum()
```

Out[10]:

```
v1               0
v2               0
Unnamed: 2    5522
Unnamed: 3    5560
Unnamed: 4    5566
dtype: int64
```

In [14]:

```
df1=df.drop(["Unnamed: 2","Unnamed: 3","Unnamed: 4"],axis=1)
```

In [15]:

```
df1
```

Out[15]:

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

5572 rows × 2 columns

In [16]:

```
df.loc[df['v1']=='spam','v1',]=0
df.loc[df['v1']=='spam','v1',]=1
```

In [17]:

```
df
```

Out[17]:

	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	0	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	0	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

5572 rows × 5 columns

In [18]:

```
df1.shape
```

Out[18]:

```
(5572, 2)
```

In [19]:

```
tfidf_vectorizer=TfidfVectorizer(max_features=5000)
x=tfidf_vectorizer.fit_transform(data['message']).toarray()
```

In [26]:

```
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import classification_report,accuracy_score
x_train,x_test,y_train,y_test=train_test_split(x,data['label'],test_size=0.2,random_state=42)


nb_classifier=MultinomialNB()
nb_classifier.fit(x_train,y_train)
y_pred=nb_classifier.predict(x_test)

print("Accuracy:",accuracy_score(y_test,y_pred))
print("\nClassification Report:\n", classification_report(y_test,y_pred))
```

```
Accuracy: 0.9739910313901345

Classification Report:
               precision    recall  f1-score   support

         ham       0.97      1.00      0.99       965
        spam       1.00      0.81      0.89       150

    accuracy                           0.97      1115
   macro avg       0.99      0.90      0.94      1115
weighted avg       0.97      0.97      0.97      1115
```

In [30]:

```
new_email="Hey, Congratulations ! you won a scratch card."
preprocessed_new_email=pre_process_text(new_email)
new_email_tfidf=tfidf_vectorizer.transform([preprocessed_new_email]).toarray()
prediction=nb_classifier.predict(new_email_tfidf)
print("Predicted Label:",prediction[0])
```

```
Predicted Label: ham
```

In [33]:

```
new_email="Hey, Congratulations ! you've won a free gift card. Click here to claim."
preprocessed_new_email=pre_process_text(new_email)
new_email_tfidf=tfidf_vectorizer.transform([preprocessed_new_email]).toarray()
prediction=nb_classifier.predict(new_email_tfidf)
print("Predicted Label:",prediction[0])
```

```
Predicted Label: spam
```

In [ ]:

```

```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js