Naive-Bayes Algorithm implementation

In [1]:
```
import pandas as pd
import numpy as np
```

importing the Dataset

In [2]:
```
data = pd.read_csv(r"https://raw.githubusercontent.com/sunnysavita10/Naive-Bayes/main/SpamClassifier-with-ML/sms_spam_data/SMSSpa
```

In [3]:
```
data.head()
```

Out[3]:

	label	messages
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...

In [4]:
```
data.shape
```

Out[4]:
```
(5572, 2)
```

In [6]:
```
data["messages"][0] #giving text data.
```

Out[6]:
```
'Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...'
```

In [7]:
```
data["messages"][50]
```

Out[7]:
```
'What you thinked about me. First time you saw me in class.'
```

In [8]:
```
data["messages"][40]
```

Out[8]:
```
'Pls go ahead with watts. I just wanted to be sure. Do have a great weekend. Abiola'
```

Data cleaning and preprocessing

In [9]:
```
import nltk # its just libraries as pandas & numpy .
```

In [10]:
```
import re #regular expression. #find out the data our text data set .
```

In [73]:
```
nltk.download("stopwords")
```

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

Out[73]:
```
True
```

In [74]:
```
#import stopwords
from nltk.corpus import stopwords
```

In [75]:
```
# import porterstemmer
from nltk.stem.porter import PorterStemmer
```

In [76]:
```
# create object this one..
ps=PorterStemmer()
```

In [77]:
```
stopwords.words('english')
```

```
 'am',
 'is',
 'are',
 'was',
 'were',
 'be',
 'been',
 'being',
 'have',
 'has',
 'had',
 'having',
 'do',
 'does',
 'did',
 'doing',
 'a',
 'an',
 'the',
 'and',
```

In [78]:
```
rev=re.sub('[^a-zA-Z]',' ',data['messages'][0])
```

In [79]:
```
rev.lower()
```

Out[79]:
```
'go until jurong point  crazy   available only in bugis n great world la e buffet    cine there got amore wat   '
```

In [80]:
```
rev=rev.split()
```

In [81]:
```
rev=' '.join(rev)
```

In [82]:
```
rev
```

Out[82]:
```
'Go until jurong point crazy Available only in bugis n great world la e buffet Cine there got amore wat'
```

In [83]:
```
corpus=[]
```

In [84]:
```
for i in range(0, len(data)):
    review = re.sub('[^a-zA-Z]', ' ', data['messages'][i])
    review = review.lower()
    review = review.split()

    review = [ps.stem(word) for word in review if not word in stopwords.words('english')]
    review = ' '.join(review)
    corpus.append(review)
```

In [85]:
```
corpus
```

Out[85]:
```
['go jurong point crazi avail bugi n great world la e buffet cine got amor wat',
 'ok lar joke wif u oni',
 'free entri wkli comp win fa cup final tkt st may text fa receiv entri question std txt rate c appli',
 'u dun say earli hor u c alreadi say',
 'nah think goe usf live around though',
 'freemsg hey darl week word back like fun still tb ok xxx std chg send rcv',
 'even brother like speak treat like aid patent',
 'per request mell mell oru minnaminungint nurungu vettam set callertun caller press copi friend callertun',
 'winner valu network custom select receivea prize reward claim call claim code kl valid hour',
 'mobil month u r entitl updat latest colour mobil camera free call mobil updat co free',
 'gonna home soon want talk stuff anymor tonight k cri enough today',
 'six chanc win cash pound txt csh send cost p day day tsandc appli repli hl info',
 'urgent week free membership prize jackpot txt word claim c www dbuk net lccltd pobox ldnw rw',
 'search right word thank breather promis wont take help grant fulfil promis wonder bless time',
 'date sunday',
 'xxxmobilemovieclub use credit click wap link next txt messag click http wap xxxmobilemovieclub com n qjkgighjjgcbl',
 'oh k watch',
 'eh u rememb spell name ye v naughti make v wet',
 'fine way u feel way gota b',
```

In [86]:
```
#Creating the Bag of Words model

from sklearn.feature_extraction.text import CountVectorizer
```

In [87]:
```
cv = CountVectorizer(max_features=2500)
```

In [88]:
```
X = cv.fit_transform(corpus).toarray()
```

In [89]:
```
#1 set of unique words
#2 finally it is creating a vectore
X.shape
```

Out[89]:
```
(5572, 2500)
```

In [90]:
```
len(X[0])
```

Out[90]:
```
2500
```

In [91]:
```
data['label']
```

Out[91]:
```
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 [92]:
```
y=pd.get_dummies(data['label'],drop_first=True)
```

In [93]:
```
X
```

Out[93]:
```
array([[0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       ...,
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

In [94]:
```
y
```

Out[94]:

	spam
0	0
1	0
2	1
3	0
4	0
...	...
5567	1
5568	0
5569	0
5570	0
5571	0

5572 rows × 1 columns

Train Test split

In [95]:
```
from sklearn.model_selection import train_test_split
```

In [96]:
```
X_train, X_test, y_train, y_test=train_test_split(X,y,test_size=0.25,random_state=10)
```

In [97]:
```
from sklearn.naive_bayes import GaussianNB
```

In [98]:
```
model=GaussianNB()
```

In [99]:
```
model.fit(X_train,y_train)
```

```
C:\Users\Balodi\AppData\Roaming\Python\Python39\site-packages\sklearn\utils\validation.py:1111: DataConversionWarning: A column
-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
  y = column_or_1d(y, warn=True)
```

Out[99]:
```
GaussianNB
GaussianNB()
```

In [100]:
```
y_pred=model.predict(X_test)
```

In [101]:
```
from sklearn.metrics import accuracy_score
```

In [102]:
```
accuracy_score(y_test,y_pred)
```

Out[102]:
```
0.8628858578607322
```

In [108]:
```
# Training model using Naive bayes classifier...
```

In [103]:
```
from sklearn.naive_bayes import MultinomialNB
```

In [104]:
```
model2=MultinomialNB()
```

In [105]:
```
model2.fit(X_train,y_train)
```

```
C:\Users\Balodi\AppData\Roaming\Python\Python39\site-packages\sklearn\utils\validation.py:1111: DataConversionWarning: A column
-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
  y = column_or_1d(y, warn=True)
```

Out[105]:
```
MultinomialNB
MultinomialNB()
```

In [106]:
```
y_pred2=model2.predict(X_test)
```

In [107]:
```
accuracy_score(y_test,y_pred2)
```

Out[107]:
```
0.9770279971284996
```

In [ ]:
```

```