spam-email-detection

April 9, 2024

IMPORTING REQUIRED LIBRARIES

[1]:
```
import numpy as np
import pandas as pd
import nltk
import re
import matplotlib.pyplot as plt
import seaborn as sns
```

IMPORTING DATASET

[2]:
```
df=pd.read_csv('/content/spam.csv')
df
```

[2]:
```
     Category                                             Message
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 ü 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 x 2 columns]
```

DATA PREPROCESSING

[3]:
```
# printing first 5 rows
df.head()
```

[3]:
```
  Category                                             Message
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…
```

```
[4]: # printing last 5 rows
     df.tail()
```

```
[4]:      Category                                            Message
     5567     spam  This is the 2nd time we have tried 2 contact u…
     5568      ham               Will ü 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
```

```
[5]: # printing datatypes
     df.dtypes
```

```
[5]: Category    object
     Message     object
     dtype: object
```

```
[6]: #finding out missing values
     df.isna().sum()
```

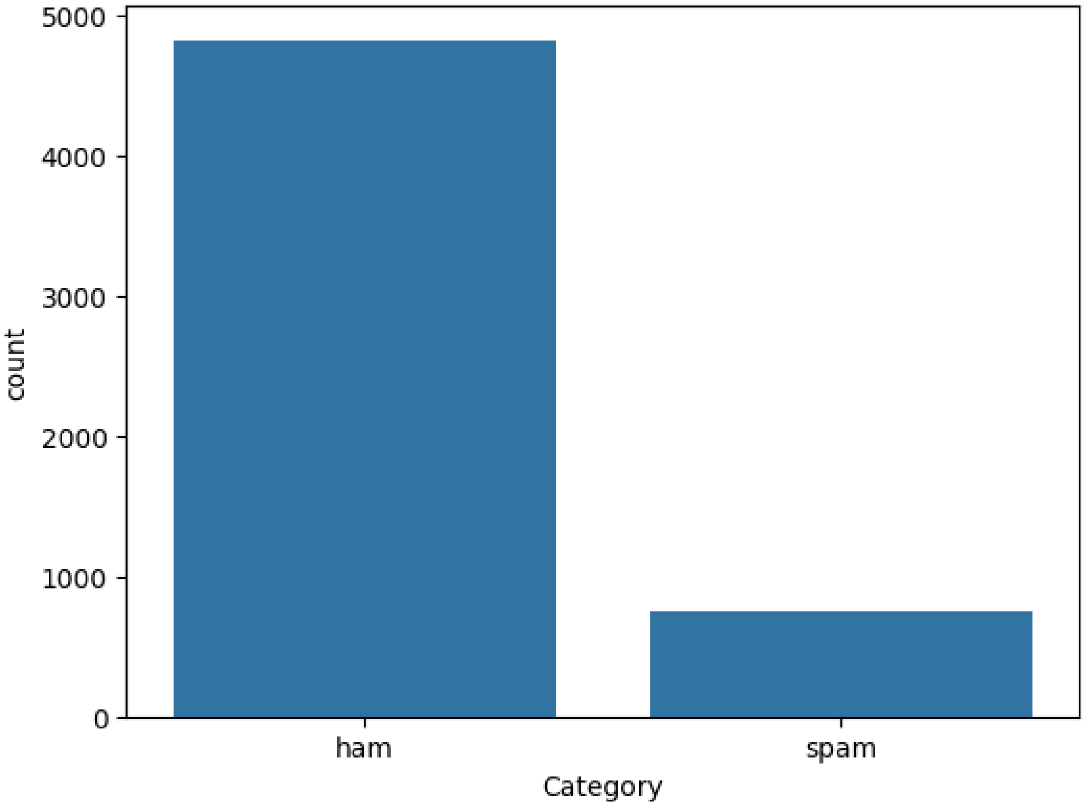
```
[6]: Category    0
     Message     0
     dtype: int64
```

```
[7]: df['Category'].value_counts()
```

```
[7]: Category
     ham     4825
     spam     747
     Name: count, dtype: int64
```

```
[8]: sns.countplot(x='Category',data=df)
```

```
[8]: <Axes: xlabel='Category', ylabel='count'>
```



[9]:
```
df['Category']=df['Category'].str.replace('ham','1')
df['Category']=df['Category'].str.replace('spam','0')
df
```

[9]:
```
     Category                                            Message
0           1  Go until jurong point, crazy.. Available only …
1           1                      Ok lar… Joking wif u oni…
2           0  Free entry in 2 a wkly comp to win FA Cup fina…
3           1  U dun say so early hor… U c already then say…
4           1  Nah I don't think he goes to usf, he lives aro…
…         …                                                …
5567        0  This is the 2nd time we have tried 2 contact u…
5568        1               Will ü b going to esplanade fr home?
5569        1  Pity, * was in mood for that. So…any other s…
5570        1  The guy did some bitching but I acted like i'd…
5571        1                         Rofl. Its true to its name

[5572 rows x 2 columns]
```

[10]:
```
df['Category']=df['Category'].astype(int)
df.dtypes
```

```
[10]: Category     int64
      Message     object
      dtype: object
```

```
[11]: df
```

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

      [5572 rows x 2 columns]
```

```
[12]: emails=df.Message
      emails
```

```
[12]: 0       Go until jurong point, crazy.. Available only …
      1                           Ok lar… Joking wif u oni…
      2       Free entry in 2 a wkly comp to win FA Cup fina…
      3       U dun say so early hor… U c already then say…
      4       Nah I don't think he goes to usf, he lives aro…
                                    …
      5567    This is the 2nd time we have tried 2 contact u…
      5568                 Will ü b going to esplanade fr home?
      5569    Pity, * was in mood for that. So…any other s…
      5570    The guy did some bitching but I acted like i'd…
      5571                           Rofl. Its true to its name
      Name: Message, Length: 5572, dtype: object
```

```
[13]: nltk.download('stopwords')
      nltk.download('punkt')
      nltk.download('wordnet')
      nltk.download('omw-1.4')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data…
[nltk_data]   Unzipping corpora/stopwords.zip.
[nltk_data] Downloading package punkt to /root/nltk_data…
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package wordnet to /root/nltk_data…
```

```
[nltk_data] Downloading package omw-1.4 to /root/nltk_data…
```

```
[13]: True
```

TOKENIZATION

```
[14]: from nltk.tokenize import TweetTokenizer
      tk=TweetTokenizer()
      emails=emails.apply(lambda x:tk.tokenize(x)).apply(lambda x:" ".join(x))
      emails
```

```
[14]: 0       Go until jurong point , crazy .. Available onl…
      1                         Ok lar … Joking wif u oni …
      2       Free entry in 2 a wkly comp to win FA Cup fina…
      3       U dun say so early hor … U c already then sa…
      4       Nah I don't think he goes to usf , he lives ar…
                                    …
      5567    This is the 2nd time we have tried 2 contact u…
      5568                Will ü b going to esplanade fr home ?
      5569    Pity , * was in mood for that . So … any oth…
      5570    The guy did some bitching but I acted like i'd…
      5571                          Rofl . Its true to its name
      Name: Message, Length: 5572, dtype: object
```

```
[15]: emails=emails.str.replace('[^A-Za-z0-9]+',' ')
      emails
```

```
[15]: 0       Go until jurong point , crazy .. Available onl…
      1                         Ok lar … Joking wif u oni …
      2       Free entry in 2 a wkly comp to win FA Cup fina…
      3       U dun say so early hor … U c already then sa…
      4       Nah I don't think he goes to usf , he lives ar…
                                    …
      5567    This is the 2nd time we have tried 2 contact u…
      5568                Will ü b going to esplanade fr home ?
      5569    Pity , * was in mood for that . So … any oth…
      5570    The guy did some bitching but I acted like i'd…
      5571                          Rofl . Its true to its name
      Name: Message, Length: 5572, dtype: object
```

```
[16]: from nltk.tokenize import word_tokenize
      emails=emails.apply(lambda x:' '.join([w for w in word_tokenize(x) if␣
       ↪len(w)>=3]))
      emails
```

```
[16]: 0       until jurong point crazy Available only bugis …
      1                              lar … Joking wif oni …
      2       Free entry wkly comp win Cup final tkts 21st M…
```

```
3              dun say early hor … already then say …
4         Nah n't think goes usf lives around here though
                              …
5567    This the 2nd time have tried contact have won …
5568                            Will going esplanade home
5569     Pity was mood for that … any other suggestions
5570    The guy did some bitching but acted like inter…
5571                                   Rofl Its true its name
Name: Message, Length: 5572, dtype: object
```

STEMMING

[17]:

```
from nltk.stem import SnowballStemmer
stemmer=SnowballStemmer('english')
emails=emails.apply(lambda x :[stemmer.stem(i.lower()) for i in tk.
 ↪tokenize(x)]).apply(lambda x:' '.join(x))
emails
```

[17]:

```
0       until jurong point crazi avail onli bugi great…
1                                lar … joke wif oni …
2       free entri wkli comp win cup final tkts 21st m…
3              dun say earli hor … alreadi then say …
4           nah n't think goe usf live around here though
                              …
5567    this the 2nd time have tri contact have won th…
5568                                    will go esplanad home
5569         piti was mood for that … ani other suggest
5570    the guy did some bitch but act like interest b…
5571                                     rofl it true it name
Name: Message, Length: 5572, dtype: object
```

REMOVING STOPWORDS

[18]:

```
from nltk.corpus import stopwords
sw=stopwords.words('english')
emails=emails.apply(lambda x:[i for i in tk.tokenize(x) if i not in sw]).
 ↪apply(lambda x:' '.join(x))
emails
```

[18]:

```
0       jurong point crazi avail onli bugi great world…
1                                lar … joke wif oni …
2       free entri wkli comp win cup final tkts 21st m…
3                   dun say earli hor … alreadi say …
4                nah n't think goe usf live around though
                              …
5567    2nd time tri contact 750 pound prize claim eas…
5568                                         go esplanad home
5569                            piti mood … ani suggest
```

```
5570    guy bitch act like interest buy someth els nex…
5571                                           rofl true name
Name: Message, Length: 5572, dtype: object
```

VECTORIZATION

```
[19]: from sklearn.feature_extraction.text import TfidfVectorizer
      vec=TfidfVectorizer()
      train_data=vec.fit_transform(emails)
```

```
[20]: train_data.shape
```

```
[20]: (5572, 7082)
```

```
[21]: y=df['Category'].values
      y
```

```
[21]: array([1, 1, 0, …, 1, 1, 1])
```

CONVERT INTO TRAINING AND TESTING DATA

```
[37]: from sklearn.model_selection import train_test_split
      x_train,x_test,y_train,y_test=train_test_split(train_data,y,test_size=0.
        ↪30,random_state=42)
      x_train
```

```
[37]: <3900x7082 sparse matrix of type '<class 'numpy.float64'>'
              with 31460 stored elements in Compressed Sparse Row format>
```

```
[38]: x_test
```

```
[38]: <1672x7082 sparse matrix of type '<class 'numpy.float64'>'
              with 13226 stored elements in Compressed Sparse Row format>
```

```
[39]: y_train
```

```
[39]: array([1, 1, 1, …, 1, 1, 1])
```

```
[40]: y_test
```

```
[40]: array([1, 1, 1, …, 1, 0, 1])
```

MODEL CRAETION

```
[41]: from sklearn.ensemble import RandomForestClassifier
      from sklearn.svm import SVC
      from sklearn.tree import DecisionTreeClassifier
      from sklearn.metrics import accuracy_score,classification_report
```

```
rfc=RandomForestClassifier(n_estimators=100,random_state=42)
svc=SVC()
tree=DecisionTreeClassifier(criterion='entropy')
```

[42]:

```
lst=[rfc,svc,tree]
for i in lst:
  print('MODEL IS',i)
  i.fit(x_train,y_train)
  y_pred=i.predict(x_test)
  print('SCORE IS',accuracy_score(y_test,y_pred))
  print('*'*80)
  print('REPORT IS',classification_report(y_test,y_pred))
```

```
MODEL IS RandomForestClassifier(random_state=42)
SCORE IS 0.9808612440191388
********************************************************************************
REPORT IS               precision    recall  f1-score   support

           0       1.00      0.86      0.92       224
           1       0.98      1.00      0.99      1448

    accuracy                           0.98      1672
   macro avg       0.99      0.93      0.96      1672
weighted avg       0.98      0.98      0.98      1672

MODEL IS SVC()
SCORE IS 0.9760765550239234
********************************************************************************
REPORT IS               precision    recall  f1-score   support

           0       0.99      0.83      0.90       224
           1       0.97      1.00      0.99      1448

    accuracy                           0.98      1672
   macro avg       0.98      0.91      0.94      1672
weighted avg       0.98      0.98      0.98      1672

MODEL IS DecisionTreeClassifier(criterion='entropy')
SCORE IS 0.9575358851674641
********************************************************************************
REPORT IS               precision    recall  f1-score   support

           0       0.83      0.86      0.84       224
           1       0.98      0.97      0.98      1448

    accuracy                           0.96      1672
   macro avg       0.90      0.92      0.91      1672
```

```
weighted avg       0.96      0.96      0.96      1672
```