spam-detection

August 30, 2024

[ ]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.metrics import roc_auc_score,f1_score,confusion_matrix
from sklearn.naive_bayes import MultinomialNB
```

[ ]:
```
df = pd.read_csv('spam.csv' , encoding='windows-1252')
```

[ ]:
```
print(df)
```

```
        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]
```

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

[ ]:
```
     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

  Unnamed: 3 Unnamed: 4
0        NaN        NaN
1        NaN        NaN
2        NaN        NaN
3        NaN        NaN
4        NaN        NaN
```

[ ]:
```
df.tail()
```

[ ]:
```
        v1                                                  v2 Unnamed: 2  \
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
5567        NaN        NaN
5568        NaN        NaN
5569        NaN        NaN
5570        NaN        NaN
5571        NaN        NaN
```

[ ]:
```
df.describe
```

[ ]:
```
<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]>
```

[ ]:
```
df = df.drop(columns=df.columns[2:5])
df.head()
```

[ ]:
```
     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…
```

[ ]:
```
df.columns = ['types','message']
df
```

[ ]:
```
     types                                               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 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 x 2 columns]
```

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

[ ]:
```
types      0
message    0
dtype: int64
```

[ ]:
```
df = df.where(pd.notnull(df), ' ')
```

[ ]:
```
df.head(10)
```

[ ]:

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

[ ]:
```
df.describe()
```

[ ]:

	types	message
count	5572	5572
unique	2	5169
top	ham	Sorry, I'll call later
freq	4825	30

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5572 entries, 0 to 5571
Data columns (total 2 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   types    5572 non-null   object
 1   message  5572 non-null   object
dtypes: object(2)
memory usage: 87.2+ KB
```

[ ]:
```
df.shape
```

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

[ ]:
```
# Replace "spam" with 0 and "ham" with 1 in the "types" column
df.loc[df["types"] == "spam", "types"] = 0
df.loc[df["types"] == "ham", "types"] = 1
```

[ ]:
```
X = df["message"]
Y = df["types"]
```

[ ]:
```
print(X)
```

```
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 Ì_ 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
```

[ ]:
```
print(Y)
```

```
0       1
1       1
2       0
3       1
4       1
       ..
5567    0
5568    1
5569    1
5570    1
5571    1
Name: types, Length: 5572, dtype: object
```

[ ]:
```
X_train,X_test,Y_train,Y_test = train_test_split(X,Y,test_size=0.
↪2,random_state=3)
```

[ ]:
```
print(X.shape)
print(X_train.shape)
print(X_test.shape)
```

```
(5572,)
(4457,)
(1115,)
```

[ ]:
```
print(Y.shape)
print(Y_train.shape)
print(Y_test.shape)
```

```
(5572,)
(4457,)
(1115,)
```

[ ]:
```
feature_extraction =␣
↪TfidfVectorizer(min_df=1,stop_words='english',lowercase='True')
```

[ ]:
```
feature_extraction = TfidfVectorizer(lowercase=True)

# Applying it to the data
X_train_features = feature_extraction.fit_transform(X_train)
X_test_features = feature_extraction.transform(X_test)
```

[ ]:
```
Y_train=Y_train.astype('int')
Y_test=Y_test.astype('int')
```

[ ]:
```
print(X_train)
```

```
3075    Mum, hope you are having a great day. Hoping t…
1787                          Yes:)sura in sun tv.:)lol.
1614    Me sef dey laugh you. Meanwhile how's my darli…
4304                 Yo come over carlos will be here soon
3266                     Ok then i come n pick u at engin?
                              …
789                          Gud mrng dear hav a nice day
968              Are you willing to go for aptitude class.
1667    So now my dad is gonna call after he gets out …
3321    Ok darlin i supose it was ok i just worry too …
1688                      Nan sonathaya soladha. Why boss?
Name: message, Length: 4457, dtype: object
```

[ ]:
```
print(X_train_features)
```

```
  (0, 741)      0.28307455118083463
  (0, 3360)     0.1327523238442287
  (0, 4108)     0.21196015023008544
  (0, 4908)     0.12973225055917514
  (0, 3042)     0.26300555749396887
  (0, 946)      0.1151770452043338
  (0, 7464)     0.19261204588580316
  (0, 4431)     0.3421657916670175
  (0, 6805)     0.17846848640014898
  (0, 6873)     0.15216101010779184
  (0, 3497)     0.28307455118083463
```

```
  (0, 2178)     0.33952544349598
  (0, 3235)     0.3869898904365042
  (0, 3365)     0.22753426247664568
  (0, 1032)     0.1361275560580212
  (0, 7720)     0.1765620792792692
  (0, 3491)     0.19174855251416806
  (0, 4655)     0.2558426236041184
  (1, 4190)     0.3725861907992424
  (1, 7099)     0.42172200036894236
  (1, 6620)     0.46707907862382136
  (1, 3646)     0.2020333473623602
  (1, 6645)     0.553594666958471
  (1, 7704)     0.3433404875792393
  (2, 954)      0.4257390912308466
  :     :
  (4455, 7402)  0.15130978849620824
  (4455, 6316)  0.16857953235415776
  (4455, 6850)  0.14840315498751144
  (4455, 1592)  0.11611242093594701
  (4455, 6991)  0.1743411711262433
  (4455, 4647)  0.1711510506728716
  (4455, 3888)  0.13162386869199988
  (4455, 3767)  0.1131472348271079
  (4455, 1615)  0.12678729795752244
  (4455, 1561)  0.12510711341007413
  (4455, 6956)  0.15407243057965578
  (4455, 847)   0.18793900087060836
  (4455, 2376)  0.1310118611150462
  (4455, 6845)  0.148803926710582
  (4455, 4933)  0.28083822596779895
  (4455, 4680)  0.11249715586710375
  (4455, 4410)  0.1081290969254776
  (4455, 3360)  0.23698431521172753
  (4455, 946)   0.10280480369551688
  (4455, 7720)  0.07879794914071371
  (4456, 6312)  0.5058318398291911
  (4456, 6334)  0.5058318398291911
  (4456, 1431)  0.4253166254875381
  (4456, 4703)  0.4655742326583645
  (4456, 7513)  0.3010227592699582
```

[ ]:
```
print(Y_train)
```

```
3075    1
1787    1
1614    1
4304    1
3266    1
```

```
       ..
789     1
968     1
1667    1
3321    1
1688    1
Name: types, Length: 4457, dtype: int64
```

[ ]:
```
model=LogisticRegression()
model.fit(X_train_features,Y_train)
```

[ ]:
```
LogisticRegression()
```

[ ]:
```
print(type(Y_train))
```

```
<class 'pandas.core.series.Series'>
```

[ ]:
```
print(np.unique(Y_train))
```

```
[0 1]
```

[ ]:
```
print(X_train_features.shape)
print(Y_train.shape)
```

```
(4457, 7777)
(4457,)
```

[ ]:
```
print(X_train_features.shape)
```

```
(4457, 7777)
```

[ ]:
```
print(np.unique(Y_train, return_counts=True))
```

```
(array([0, 1]), array([ 592, 3865]))
```

[ ]:
```
prediction_train=model.predict(X_train_features)
accuracy_train=accuracy_score(Y_train,prediction_train)
```

[ ]:
```
print("Accuracy:",accuracy_train)
```

```
Accuracy: 0.9739735247924612
```

[ ]:
```
prediction_test=model.predict(X_test_features)
accuracy_test=accuracy_score(Y_test,prediction_test)
```

[ ]:
```
print("accuracy:",accuracy_test)
```

```
accuracy: 0.9757847533632287
```

[ ]:
```
sample_input=["congrats,you have won a free ticket to U.K"]
input_data_features=feature_extraction.transform(sample_input)
prediction=model.predict(input_data_features)
print(prediction)

if(prediction)[0]==1:
  print("Ham mail")
else:
  print("Spam mail")
```

```
[0]
Spam mail
```

[ ]:
```
sample_input2=["Meeting at 10:am . see you there!"]
input_data2_features=feature_extraction.transform(sample_input2)
prediction2=model.predict(input_data2_features)
print(prediction2)
if(prediction2)[0]==1:
  print("Ham mail")
else:
  print("Spam mail")
```

```
[1]
Ham mail
```