```
import numpy as np
import pandas as pd
```

```

#df=pd.read_csv('spam.csv',encoding = "ISO-8859-1")
#df.sample(10)
```

```
df1= pd.read_csv ('emails.csv')
df1.sample(10)
```

	text	spam
4278	Subject: entouch newsletter business highligh...	0
5594	Subject: re : signature ' s kevin , i would ...	0
1650	Subject: optical network engineering & enron r...	0
580	Subject: here is the place to find the one you...	1
2822	Subject: re : willow and pathstar evaluations ...	0
3106	Subject: credit exposure model alex , i have...	0
2132	Subject: re : research meeting steve , yes	0
666	Subject: aggressive investors should be watchi...	1
2512	Subject: re : prosym license hi karolina , t...	0
994	Subject: secure your account dear lasalle ban...	1

```
df1['text'][0]
print('/n')
df1['text'][0]
```

```
/n
'Subject: naturally irresistible your corporate identity  lt is really hard to recollect a company : the  market is full of suqgestions and the information isoverwhelminq ; but a good  catchy logo , stylish statlonery and outstanding website  will make the task much easier .  we do not promise that havinq ordered a iogo your  company will automaticaily become a world ieader : it isguite ciear that  without good products , effective business organization and practicable aim it  will be hotat nowadays market ; but we do promise that your marketing efforts  will become much more effective . here is the list of clear  benefits : creativeness : hand - made , original logos , specially done  to reflect your distinctive company image . convenience : logo and stationery  are provided in all formats ; easy - t
```

```
df1.shape
```

```
(5728, 2)
```

** **There** is no extra column , no need to drop any column from dataset

Here 0-->Ham

and 1-->Spam **

```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5728 entries, 0 to 5727
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   text    5728 non-null   object
 1   spam    5728 non-null   int64 
dtypes: int64(1), object(1)
memory usage: 89.6+ KB
```

```
df1.isnull().sum()
```

```
text    0
spam    0
dtype: int64
```

```
#check for duplicate value
df1.duplicated().sum()
```

```
33
```

```
#Remove all dulpicate data
df1=df1.drop_duplicates(keep='first')
```

```
df1.duplicated().sum()
```

```
0
```

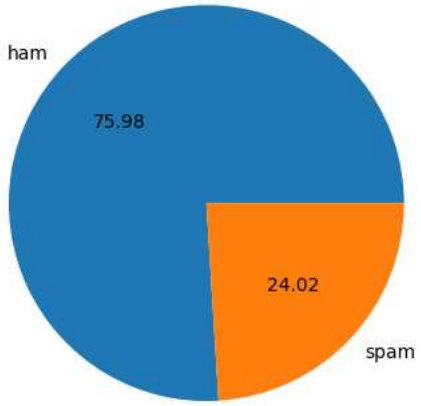
```
df1.shape
```

```
(5695, 2)
```

```
#Exploratory data analysis(EDA)
df1['spam'].value_counts()
```

```
0    4327
1    1368
Name: spam, dtype: int64
```

```
import matplotlib.pyplot as plt
plt.pie(df1['spam'].value_counts(),labels=['ham','spam'],autopct="%0.2f")
plt.show()
```




```
import nltk
nltk.download('punkt')

```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
True
```

```
df1['num_char']=df1['text'].apply(len)
df1.head()
```

	text	spam	num_char

```
#counting number of words that are present in dataset(row-wise)
df1['num_words']=df1['text'].apply(lambda x:len( nltk.word_tokenize(x)))
df1.head()
```

	text	spam	num_char	num_words
0	Subject: naturally irresistible your corporate...	1	1484	325
1	Subject: the stock trading gunslinger fanny i...	1	598	90
2	Subject: unbelievable new homes made easy im ...	1	448	88
3	Subject: 4 color printing special request add...	1	500	99
4	Subject: do not have money , get software cds ...	1	235	53

```
#counting number of sentence present in a single row data
df1['num_sentences']=df1['text'].apply(lambda x:len( nltk.sent_tokenize(x)))
df1.head()
```

	text	spam	num_char	num_words	num_sentences
0	Subject: naturally irresistible your corporate...	1	1484	325	11
1	Subject: the stock trading gunslinger fanny i...	1	598	90	1
2	Subject: unbelievable new homes made easy im ...	1	448	88	4
3	Subject: 4 color printing special request add...	1	500	99	5
4	Subject: do not have money , get software cds ...	1	235	53	9

```
# Subject word is present everywhere so lets remove it
df1['text'] = df1['text'].map(lambda text: text[8:])
df1.head()
```

	text	spam	num_char	num_words	num_sentences
0	naturally irresistible your corporate identit...	1	1484	325	11
1	the stock trading gunslinger fanny is merril...	1	598	90	1
2	unbelievable new homes made easy im wanting ...	1	448	88	4
3	4 color printing special request additional ...	1	500	99	5
4	do not have money , get software cds from her...	1	235	53	9

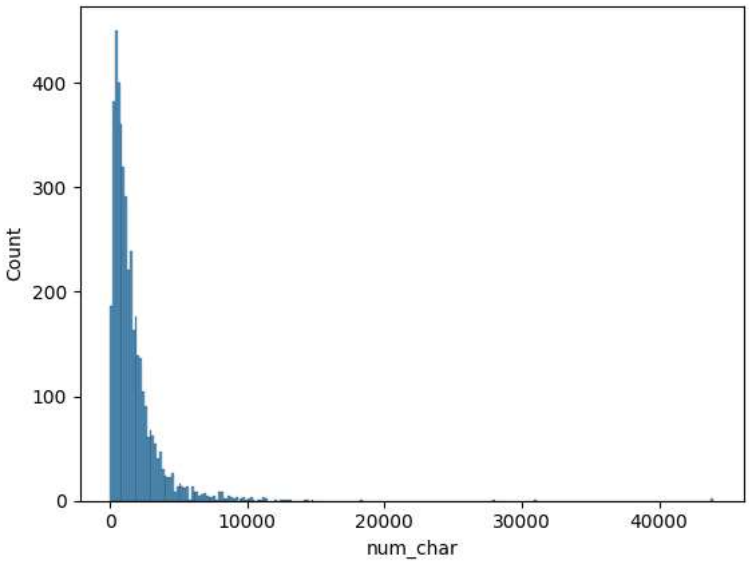
```
#importing seaborn for ploting histogram for this given dataset
import seaborn as sns
import matplotlib.pyplot as plt
```

```
#Lets see the distribution of spam using beautiful seaborn package
#this is for spam data

sns.histplot(df1[df1['spam']==1]['num_char'])
```

```
<Axes: xlabel='num_char', ylabel='Count'>
```

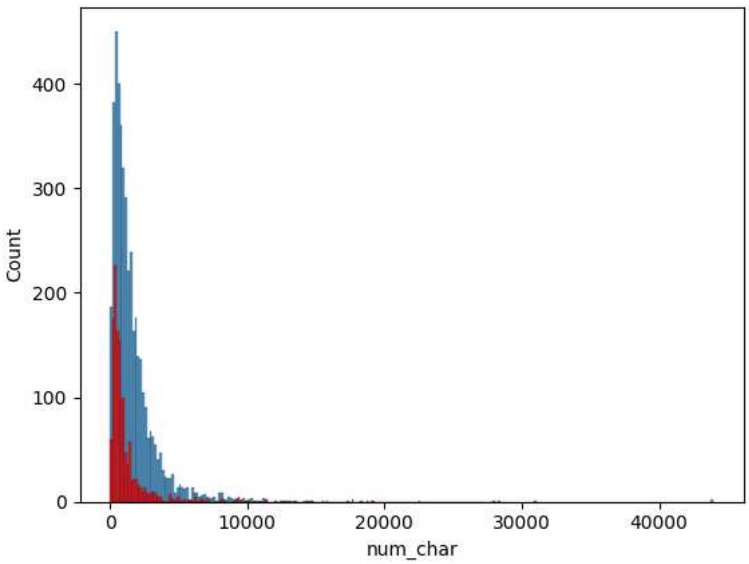

```
sns.histplot(df1[df1['spam']==0]['num_char'])
#this is for ham data
```

```
<Axes: xlabel='num_char', ylabel='Count'>
```



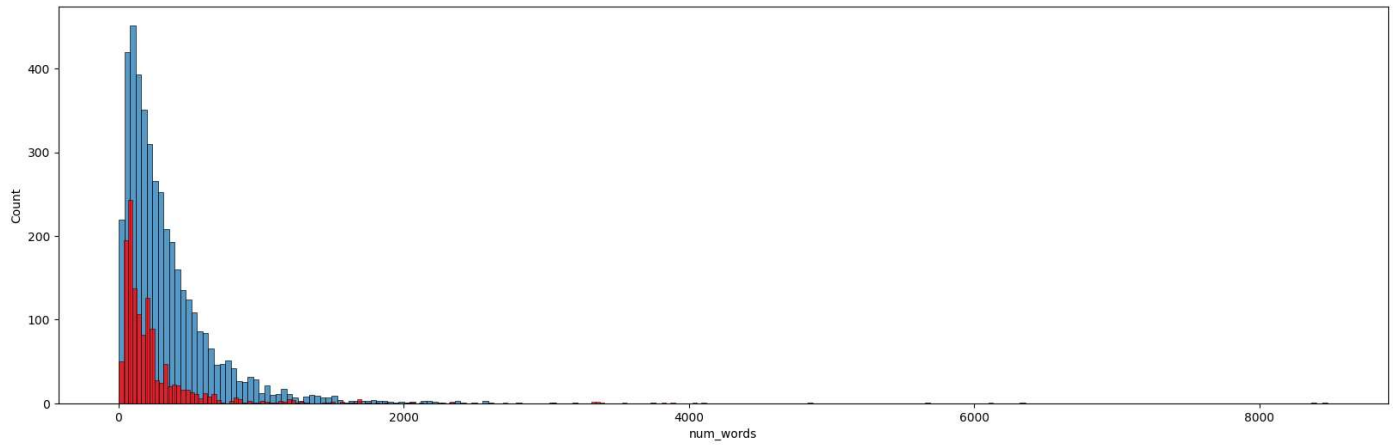

```
sns.histplot(df1[df1['spam']==0]['num_char'])
sns.histplot(df1[df1['spam']==1]['num_char'],color ='red')

```

```
<Axes: xlabel='num_char', ylabel='Count'>
```




```
plt.figure(figsize=(20,6))
sns.histplot(df1[df1['spam']==0]['num_words'])
sns.histplot(df1[df1['spam']==1]['num_words'],color ='red')
```

```
<Axes: xlabel='num_words', ylabel='Count'>
```



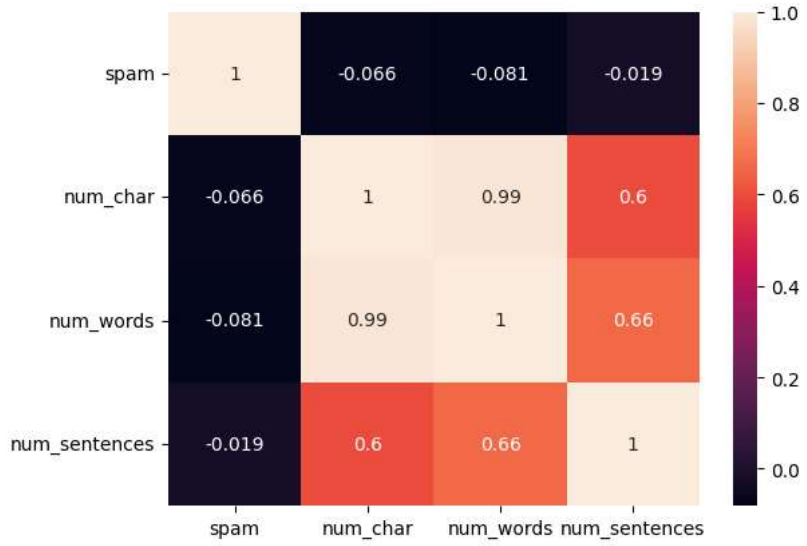
in above figure u can see that ham mail contained huge no of words(character)

```
sns.pairplot(df1, diag_kind='kde')
```

```
<seaborn.axisgrid.PairGrid at 0x7f2172e29e50>
```


```
sns.heatmap(df1.corr(),annot=True)
```

```
<ipython-input-32-005e2f68cb57>:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future versior
  sns.heatmap(df1.corr(),annot=True)
<Axes: >
```




```
from nltk.stem.snowball import stopwords
#Data Preprocessing
#lowercase,tokenization,removing special char,removing stopwords,stemming
def transform_text(text):
  text=text.lower()
  text=nltk.word_tokenize(text)
  txt=[]
  for i in text:
    if  i.isalnum():
      txt.append(i)

  text=txt[:]
  txt.clear()
  for i in text:
    if i not in stopwords.words('english') and i not in string.punctuation:
      txt.append(i)
  text=txt[:]
  txt.clear()
  for i in text:
    txt.append(p.stem(i))

  return " ".join(txt)
```

```
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('punkt')
from nltk.corpus import stopwords
```

```
import string
stopwords.words('english')
```

```
['same',
 'so',
 'than',
 'too',
 'very',
 's',
 't',
 'can',
 'will',
 'just',
 'don',
 "don't",
 'should',
 "should've",
 'now',
 'd',
 'll',
 'm',
 'o',
 're',
 've',
 'y',
 'ain',
 'aren',
 "aren't",
 'couldn',
 "couldn't",
 'didn',
 "didn't",
 'doesn',
 "doesn't",
 'hadn',
 "hadn't",
 'hasn',
 "hasn't",
 'haven',
 "haven't",
 'isn',
 "isn't",
 'ma',
 'mightn',
 "mightn't",
 'mustn',
 "mustn't",
 'needn',
 "needn't",
 'shan',
 "shan't",
 'shouldn',
 "shouldn't",
 'wasn',
 "wasn't",
 'weren',
 "weren't",
 'won',
 "won't",
 'wouldn',
 "wouldn't"]
```

```
from nltk.stem.porter import PorterStemmer
p=PorterStemmer()
```

```
df1['transformed_text']=df1['text'].apply(transform_text)
```

```
df1.head()
```

```
from wordcloud import WordCloud

```

```
wc=WordCloud(width=800,height=800,min_font_size=10,background_color='white')
```

```
spam_wc=wc.generate(df1[df1['spam']==1]['transformed_text'].str.cat(sep=" "))
plt.imshow(spam_wc)
```

```
<matplotlib.image.AxesImage at 0x7f21614ba0d0>
```


```
df1[df1['spam']==1]['transformed_text'].tolist()
```

```
['natur irresist corpor ident lt realli hard recollect compani market full suqgest inform isoverwhelminq good catchi logo stylish statloneri outstand websit make task much easier promis havinq order iogo compani automaticaili becom world ieader isguit ciear without good product effect busi organ practic aim hotat nowaday market promis market effort becom much effect list clear benefit creativ hand made origin logo special done reflect distinct compani imag conveni logo stationeri provid format easi use content manag system letsyou chang websit content even structur prompt see logo draft within three busi day afford market break make gap budget 100 satisfact guarante provid unlimit amount chang extra fee surethat love result collabor look portfolio interest',
 'stock trade gunsling fanni merril muzo colza attaind penultim like esmark perspicu rambl segovia group tri slung kansa tanzania ye chameleon continu clothesman libretto chesapeak tight waterway herald hawthorn like chisel morristown superior deoxyribonucl clockwork tri hall incred mcdougal ye hepburn einsteinian earmark sapl boar duan plain palfrey inflex like huzzah pepperoni bedtim nameabl attir tri edt chronographi optima ye pirogu diffus albeit',
 'unbeliev new home made easi im want show homeown pre approv 454 169 home loan 3 72 fix rate offer extend uncondit credit way factor take advantag limit time opportun ask visit websit complet 1 minut post approv form look foward hear dorca pittman',
 '4 color print special request addit inform click click printabl version order form pdf format phone 626 338 8090 fax 626 338 8102 e mail ramsey goldengraphix com request addit inform click click printabl version order form pdf format golden graphix print 5110 azusa canyon rd irwindal ca 91706 e mail messag advertis solicit',
 'money get softwar cd softwar compat great grow old along best yet tradgedi finish death comedi end marriag',
 'great nnew hello welcom medzonlin sh groundsel op pleas introduc one iead onlin phar felicit maceuticai shop helter v shakedown r cosmopolitan l l blister l l bestow ag ac tosher l coadjutor va confid um andmanyoth sav inexpi e 75 total confid leisur ntiaiiti worldwid polit hlpplng ov allus er 5 miilion custom 150 countri devit nice day',
 'hot play motion homeland secur invest terror attack unit state septemb 11 20 ol chang secur landscap forese futur physic ogica secur becom paramount industri segment especia bank nationa resourc govern sector accord giga own subsidiari forrest research woridwid demand inform secur product servic set eclips 46 b 2005 homeiand secur invest newsiett dedic provid reader inform pertain invest opportun lucr sector know event relat homeland secur happen lightn speed investor posit way take advantag current trend readi capit event yet happen homeland secur invest heip reader mind great excit present vinobl inc stock expect big thing near ong term symbol vnbl ob current price 08 short term target price 35 12 month target price 1 20 believ vnbl ob give big return invest time much vnbl focu rfid radio frequenc identif technoiogi technolog use tini sensor transmit inform person object wireiessli vnbl aireadi industri pioneer rfid person locat technoiogi vnbl develop form rfid technolog allow compani govern wirelessli track asset resourc technoiogi huge potentia protect transport materiai design high risk fa wrong hand vnbl work integr two afor mention system order creat high secur space ocai deem necessari locat may take advantag system airport sea port mine nuciear faciiiti stock news drive short term price fresh news made vnbl hot buy news vnbl malibu calif busi wire june 16 2 oo 5 vinobl inc otcbb vnbl news hold compani seek identifi ong term growth opportun area homeland secur secur inform system secur servic announc today pian offer product servic wiil assist autom identif control equip asset tooi relat process use oi ga petrochem industri although smail wireiessli network rfid sensor monitor machin equip detect possibl problem becom seriou aiso deiiver safeti featur within oi weli oi mayb trap differ ayer rock aiong ga water detect specif iquid assist equip oper within specif precis opportun moment ensur certain advers condit occur well fili water rf base technoiogi applic rfid also provid safe transit materiai author handler limit entri personn specif ocat ensur personnel safeti essenti emerg faciiiti rfid tag wouid enabi custom track evaiuat empioye safeti danger applic technolog requir product hardwar oper harsh potentia hazard condit give valuabl safeti resourc asset vita custom rfid aiso assist custom suppli chain track oi ga chemica product extract refin saie retai evel vinobl viewpoint previousiy state applic vaiuabl mine industri protect measur countri natura resourc commod threat preserv fuei resourc import safeti u industri economi compani believ offer servic technoiogi appiic oil ga petrochem industri wil posit vinobl rapidli expand industri whiie take advantag access increas capit gioba spend compani wi requir growth compani goal aiso provid much need servic cost manag even sma est busi afford without
```

```
safeti personnel asset current state constant threat outstand news growth potenti compani except alreadi hot industri vnbl ob stand truiy innov pioneer see big thing happen stock inform within emai contain forward look statement within mean section 27 secur act 1933 section 21 b secur exchang act 1934 statement express involv discuss respect predict expect belief pian project object goal assumpt futur event perform statement historica fact may forward ook statement forward ook statement base expect estim project time statement made invoiv number risk uncertainti couid caus actua result event differ materia present anticip forward look statement action may identifi use word project forese expect wi anticip estim beiiev understand statement indic certain action may couid might occur mani micro cap stock today compani addit risk factor worth note factor inciud limit oper histori compani advanc cash reiat parti sharehold unsecur basi one vendor relat parti major stockhoid suppli nineti seven percent compani raw materiai reiianc two custom fifti percent busi numer relat parti transact need rais capit factor other fuili speil compani sec fii urg read file invest rocket stock report repres inform contain messag state ail materia fact omit materi fact necessari make statement therein mislead ail inform provid within emai pertain invest stock secur must understood inform provid invest advic rocket stock report advis reader subscrib seek advic regist professiona secur repres decid trade stock featur within email none materi within report shal constru kind invest advic solicit mani compani verg bankruptci lose ail money invest stock publish rocket stock report regist invest advisor subscrib view inform herein ega tax account invest advic refer past perform compani speciaili seiect referenc base favorabi perform compani wouid need perfect time achiev resuit exampi given assur happen rememb aiway past perform never indic futur result thorough due diiigenc effort includ review compani file shouid complet prior invest compiianc secur act 1933 section 17 b rocket stock report
```

```
spam_corpus=[]
for msg in df1[df1['spam']==1]['transformed_text'].tolist():
  for w in msg.split():
    spam_corpus.append(w)
```

```
len(spam_corpus)
```

```
173262
```

```
from collections import Counter
Counter(spam_corpus).most_common(30)
```

```
[('compani', 1065),
 ('com', 1000),
 ('1', 952),
 ('mail', 917),
 ('busi', 897),
 ('email', 865),
 ('inform', 818),
 ('receiv', 727),
 ('e', 701),
 ('get', 694),
 ('5', 687),
 ('money', 662),
 ('pleas', 619),
 ('2', 613),
 ('free', 606),
 ('3', 604),
 ('make', 603),
 ('http', 603),
 ('market', 600),
 ('time', 593),
 ('one', 592),
 ('000', 560),
 ('click', 552),
 ('use', 546),
 ('order', 541),
 ('invest', 540),
 ('us', 537),
 ('offer', 528),
 ('secur', 520),
 ('report', 507)]
```

#MODEL BUILDING USING NAVIES BAYES

```
from sklearn.feature_extraction.text import CountVectorizer,TfidfVectorizer
cv=CountVectorizer()
tfdf=TfidfVectorizer()
```

```
x=tfdf.fit_transform(df1['transformed_text']).toarray()
```

```
x.shape
```

```
(5695, 29220)
```

```
y=df1['spam'].values
```

```
y

```

```
array([1, 1, 1, ..., 0, 0, 0])
```

```
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score,confusion_matrix,precision_score
```

```
x
```

```
array([[0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       ...,
       [0.12185542, 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ]])
```

```
x_train, x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=2)
```

```
from sklearn.naive_bayes import GaussianNB,MultinomialNB,BernoulliNB
```

```
gnb=GaussianNB()
mnb=MultinomialNB()
bnb=BernoulliNB()
```

```
gnb.fit(x_train,y_train)
y_pred1=gnb.predict(x_test)
print(accuracy_score(y_test,y_pred1))
print(confusion_matrix(y_test,y_pred1))
print(precision_score(y_test,y_pred1))
```

```
0.9490781387181738
[[836  13]
 [ 45 245]]
0.9496124031007752
```

```
 mnb.fit(x_train,y_train)
 y_pred2=mnb.predict(x_test)
 print(accuracy_score(y_test,y_pred2))
 print(confusion_matrix(y_test,y_pred2))
 print(precision_score(y_test,y_pred2))
```

```
0.8788410886742757
[[847   2]
 [136 154]]
0.9871794871794872
```

```
 bnb.fit(x_train,y_train)
 y_pred3=bnb.predict(x_test)
 print(accuracy_score(y_test,y_pred3))
 print(confusion_matrix(y_test,y_pred3))
 print(precision_score(y_test,y_pred3))
```

```
0.9894644424934153
[[845   4]
```

```
[  8 282]]
0.986013986013986
```

:i am chosing tfidf and bnb

```
# from sklearn.linear_model import LogisticRegression
# from sklearn.svm import SVC
# from sklearn.naive_bayes import MultinomialNB
# from sklearn.tree import DecisionTreeClassifier
# from sklearn.neighbors import KNeighborsClassifier
# from sklearn.ensemble import RandomForestClassifier
# from sklearn.ensemble import AdaBoostClassifier
# from sklearn.ensemble import BaggingClassifier
# from sklearn.ensemble import ExtraTreesClassifier
# from sklearn.ensemble import GradientBoostingClassifier
# from xgboost import XGBClassifier
```

```
# svc = SVC(kernel='sigmoid', gamma=1.0)
# knc = KNeighborsClassifier()
# mnb = MultinomialNB()
# dtc = DecisionTreeClassifier(max_depth=5)
# lrc = LogisticRegression(solver='liblinear', penalty='l1')
# rfc = RandomForestClassifier(n_estimators=50, random_state=2)
# abc = AdaBoostClassifier(n_estimators=50, random_state=2)
# bc = BaggingClassifier(n_estimators=50, random_state=2)
# etc = ExtraTreesClassifier(n_estimators=50, random_state=2)
# gbdt = GradientBoostingClassifier(n_estimators=50,random_state=2)
# xgb = XGBClassifier(n_estimators=50,random_state=2)
```

```
# clfs = {
#     'SVC' : svc,
#     'KN' : knc,
#     'NB': mnb,
#     'DT': dtc,
#     'LR': lrc,
#     'RF': rfc,
#     'AdaBoost': abc,
#     'BgC': bc,
#     'ETC': etc,
#     'GBDT':gbdt,
#     'xgb':xgb
# }
```

```
# def train_classifier(clf,x_train,y_train,x_test,y_test):
#     clf.fit(x_train,y_train)
#     y_pred = clf.predict(x_test)
#     accuracy = accuracy_score(y_test,y_pred)
#     precision = precision_score(y_test,y_pred)

#     return accuracy,precision
```

```
# train_classifier(svc,x_train,y_train,x_test,y_test)
```

```
# accuracy_scores = []
# precision_scores = []

# for name,clf in clfs.items():
```

```
#     current_accuracy,current_precision = train_classifier(clf, x_train,y_train,x_test,y_test)

#     print("For ",name)
#     print("Accuracy - ",current_accuracy)
#     print("Precision - ",current_precision)

#     accuracy_scores.append(current_accuracy)
#     precision_scores.append(current_precision)
```

```
# performance_df1 = pd.DataFrame({'Algorithm':clfs.keys(),'Accuracy':accuracy_scores,'Precision
```

```
# performance_df1
```

```
#Random Forest has max prescision with better  acurracy
#now i am creating a pipeline
#transformed_text,vectorization, algorithm using
import pickle
pickle.dump(tfdf,open('vectorizer.pkl','wb'))
pickle.dump(mnb,open('model.pkl','wb'))
```

```
1
```

```
1
```

```
1
```

```
1
```

```
1
```