Problem statement: For FlipItNews using news articles extracted from companys' internal database categorize them into categories such as politics, technology, sports, business and entertainment based on their content. Using NLP create and compare atleast 3 different models.

In [1]:

```python
# Importing libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from string import punctuation
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer

from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split

from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorize
from sklearn.naive_bayes import GaussianNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix
```

```
C:\Users\vidya\anaconda3\lib\site-packages\scipy\__init__.py:146: UserWarn
ing: A NumPy version >=1.16.5 and <1.23.0 is required for this version of
SciPy (detected version 1.26.4
  warnings.warn(f"A NumPy version >={np_minversion} and <{np_maxversion}"
```

In [2]:

```python
# Load data
finews_d = pd.read_csv('flipitnews-data.csv')

# First few rows of the dataset
finews_d.head()
```

Out[2]:

	Category	Article
0	Technology	tv future in the hands of viewers with home th...
1	Business	worldcom boss left books alone former worldc...
2	Sports	tigers wary of farrell gamble leicester say ...
3	Sports	yeading face newcastle in fa cup premiership s...
4	Entertainment	ocean s twelve raids box office ocean s twelve...

In [3]:
```
# Last few rows of the dataset
finews_d.tail()
```

Out[3]:

	Category	Article
2220	Business	cars pull down us retail figures us retail sal...
2221	Politics	kilroy unveils immigration policy ex-chatshow ...
2222	Entertainment	rem announce new glasgow concert us band rem h...
2223	Politics	how political squabbles snowball it s become c...
2224	Sports	souness delight at euro progress boss graeme s...

In [4]:
```
# Dimensions
finews_d.ndim
```

Out[4]:
```
2
```

In [5]:
```
# Shape
finews_d.shape
```

Out[5]:
```
(2225, 2)
```

In [6]:
```
# Size
finews_d.size
```

Out[6]:
```
4450
```

In [7]:
```
# Columns
finews_d.columns
```

Out[7]:
```
Index(['Category', 'Article'], dtype='object')
```

In [8]:
```
# Index
finews_d.index
```

Out[8]:
```
RangeIndex(start=0, stop=2225, step=1)
```

In [9]:
```
# Info
finews_d.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2225 entries, 0 to 2224
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   Category  2225 non-null   object
 1   Article   2225 non-null   object
dtypes: object(2)
memory usage: 34.9+ KB
```

In [10]:
```
# Describe
finews_d.describe()
```

Out[10]:

	Category	Article
count	2225	2225
unique	5	2126
top	Sports	kennedy questions trust of blair lib dem leade...
freq	511	2

In [11]:
```
finews_d.nunique()
```

Out[11]:
```
Category       5
Article     2126
dtype: int64
```

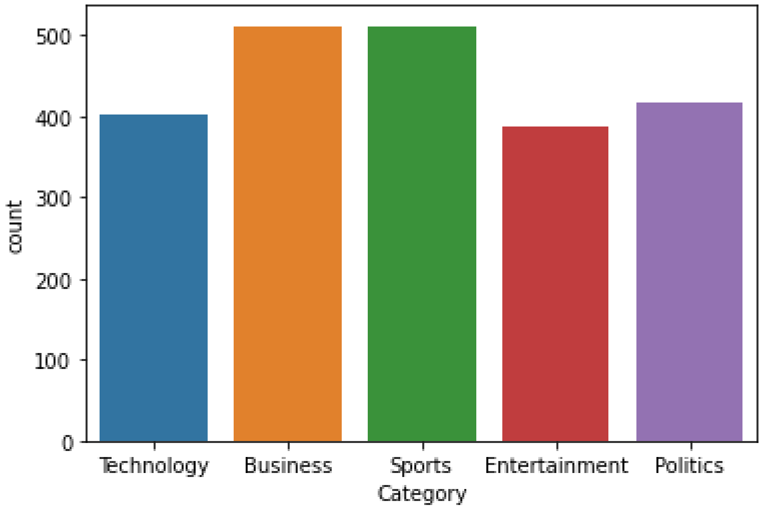
In [12]:
```
finews_d['Category'].unique()
```

Out[12]:
```
array(['Technology', 'Business', 'Sports', 'Entertainment', 'Politics'],
      dtype=object)
```

In [13]:
```
finews_d['Category'].value_counts()
```

Out[13]:
```
Sports           511
Business         510
Politics         417
Technology       401
Entertainment    386
Name: Category, dtype: int64
```

In [14]:
```
# Category
sns.countplot(data=finews_d, x='Category')
plt.show()
```



Sports have highest count

In [15]:

```
# Null value detection
finews_d.isna().sum()
```

Out[15]:

```
Category    0
Article     0
dtype: int64
```

In [16]:

```
# Detecting duplicates
finews_d[finews_d.duplicated()]
```

Out[16]:

	Category	Article
85	Politics	hague given up his pm ambition former conser...
301	Politics	fox attacks blair s tory lies tony blair lie...
496	Technology	microsoft gets the blogging bug software giant...
543	Business	economy strong in election year uk businesse...
582	Entertainment	ray dvd beats box office takings oscar-nominat...
...	...	...
2206	Politics	kennedy questions trust of blair lib dem leade...
2207	Technology	california sets fines for spyware the makers o...
2213	Technology	progress on new internet domains by early 2005...
2215	Technology	junk e-mails on relentless rise spam traffic i...
2217	Technology	rings of steel combat net attacks gambling is ...

99 rows × 2 columns

In [17]:

```
# Dropping duplicates
finews_d.drop_duplicates(keep='first', inplace=True)
```

Function to process textual data

In [18]:

```
Stp_wrd = stopwords.words('english')
def preproc_st(acl):
    # Removing non-letters
    for x in punctuation:
        if x in acl:
            acl = acl.replace(x, '')

    # Removing stopwords
    acl_n = []
    acl_l = acl.split()
    for x2 in acl_l:
        if x2 not in Stp_wrd:
            acl_n.append(x2)
    acl_n = " ".join(acl_n)
    # Word Tokenization
    acl = word_tokenize(acl_n)

    # Lemmitization
    acl_lemmatize = WordNetLemmatizer()
    acl_lem = []
    for x3 in acl:
        lm = acl_lemmatize.lemmatize(x3, pos='v')
        acl_lem.append(lm)
    acl_lem = " ".join(acl_lem)
    return acl_lem
```

In [19]:

```
# Creating column with preprocessed textual data
finews_d['nw_Article'] = finews_d['Article'].apply(preproc_st)
```

News article before and after the processing

```
In [20]: # Before preprocessing
         finews_d['Article'][0]
```

```
Out[20]: 'tv future in the hands of viewers with home theatre systems  plasma high-
          definition tvs  and digital video recorders moving into the living room  t
          he way people watch tv will be radically different in five years  time.  t
          hat is according to an expert panel which gathered at the annual consumer
          electronics show in las vegas to discuss how these new technologies will i
          mpact one of our favourite pastimes. with the us leading the trend  progra
          mmes and other content will be delivered to viewers via home networks  thr
          ough cable  satellite  telecoms companies  and broadband service providers
          to front rooms and portable devices.  one of the most talked-about technol
          ogies of ces has been digital and personal video recorders (dvr and pvr).
          these set-top boxes  like the us s tivo and the uk s sky+ system  allow pe
          ople to record  store  play  pause and forward wind tv programmes when the
          y want.  essentially  the technology allows for much more personalised tv.
          they are also being built-in to high-definition tv sets  which are big bus
          iness in japan and the us  but slower to take off in europe because of the
          lack of high-definition programming. not only can people forward wind thro
          ugh adverts  they can also forget about abiding by network and channel sch
          edules  putting together their own a-la-carte entertainment. but some us n
          etworks and cable and satellite companies are worried about what it means
          for them in terms of advertising revenues as well as  brand identity  and
          viewer loyalty to channels. although the us leads in this technology at th
          e moment  it is also a concern that is being raised in europe  particularl
          y with the growing uptake of services like sky+.  what happens here today
          we will see in nine months to a years  time in the uk   adam hume  the bbc
          broadcast s futurologist told the bbc news website. for the likes of the b
          bc  there are no issues of lost advertising revenue yet. it is a more pres
          sing issue at the moment for commercial uk broadcasters  but brand loyalty
          is important for everyone.  we will be talking more about content brands r
          ather than network brands   said tim hanlon  from brand communications fir
          m starcom mediavest.  the reality is that with broadband connections  anyb
          ody can be the producer of content.  he added:  the challenge now is that
          it is hard to promote a programme with so much choice.   what this means
          said stacey jolna  senior vice president of tv guide tv group  is that the
          way people find the content they want to watch has to be simplified for tv
          viewers. it means that networks  in us terms  or channels could take a lea
          f out of google s book and be the search engine of the future  instead of
          the scheduler to help people find what they want to watch. this kind of ch
          annel model might work for the younger ipod generation which is used to ta
          king control of their gadgets and what they play on them. but it might not
          suit everyone  the panel recognised. older generations are more comfortabl
          e with familiar schedules and channel brands because they know what they a
          re getting. they perhaps do not want so much of the choice put into their
          hands  mr hanlon suggested.  on the other end  you have the kids just out
          of diapers who are pushing buttons already - everything is possible and av
          ailable to them   said mr hanlon.  ultimately  the consumer will tell the
          market they want.   of the 50 000 new gadgets and technologies being showc
          ased at ces  many of them are about enhancing the tv-watching experience.
          high-definition tv sets are everywhere and many new models of lcd (liquid
          crystal display) tvs have been launched with dvr capability built into the
          m  instead of being external boxes. one such example launched at the show
          is humax s 26-inch lcd tv with an 80-hour tivo dvr and dvd recorder. one o
          f the us s biggest satellite tv companies  directtv  has even launched its
          own branded dvr at the show with 100-hours of recording capability  instan
          t replay  and a search function. the set can pause and rewind tv for up to
          90 hours. and microsoft chief bill gates announced in his pre-show keynote
          speech a partnership with tivo  called tivotogo  which means people can pl
          ay recorded programmes on windows pcs and mobile devices. all these reflec
          t the increasing trend of freeing up multimedia so that people can watch w
          hat they want  when they want.'
```

In [21]:

```
# After preprocessing
finews_d['nw_Article'][0]
```

Out[21]:

```
'tv future hand viewers home theatre systems plasma highdefinition tvs dig
ital video recorders move live room way people watch tv radically differen
t five years time accord expert panel gather annual consumer electronics s
how las vegas discuss new technologies impact one favourite pastimes us le
ad trend program content deliver viewers via home network cable satellite
telecoms company broadband service providers front room portable devices o
ne talkedabout technologies ces digital personal video recorders dvr pvr s
ettop box like us tivo uk sky system allow people record store play pause
forward wind tv program want essentially technology allow much personalise
tv also builtin highdefinition tv set big business japan us slower take eu
rope lack highdefinition program people forward wind advert also forget ab
ide network channel schedule put together alacarte entertainment us networ
k cable satellite company worry mean term advertise revenues well brand id
entity viewer loyalty channel although us lead technology moment also conc
ern raise europe particularly grow uptake service like sky happen today se
e nine months years time uk adam hume bbc broadcast futurologist tell bbc
news website like bbc issue lose advertise revenue yet press issue moment
commercial uk broadcasters brand loyalty important everyone talk content b
rand rather network brand say tim hanlon brand communications firm starcom
mediavest reality broadband connections anybody producer content add chall
enge hard promote programme much choice mean say stacey jolna senior vice
president tv guide tv group way people find content want watch simplify tv
viewers mean network us term channel could take leaf google book search en
gine future instead scheduler help people find want watch kind channel mod
el might work younger ipod generation use take control gadgets play might
suit everyone panel recognise older generations comfortable familiar sched
ule channel brand know get perhaps want much choice put hand mr hanlon sug
gest end kid diapers push button already everything possible available say
mr hanlon ultimately consumer tell market want 50 000 new gadgets technolo
gies showcased ces many enhance tvwatching experience highdefinition tv se
t everywhere many new model lcd liquid crystal display tvs launch dvr capa
bility build instead external box one example launch show humax 26inch lcd
tv 80hour tivo dvr dvd recorder one us biggest satellite tv company direct
tv even launch brand dvr show 100hours record capability instant replay se
arch function set pause rewind tv 90 hours microsoft chief bill gate annou
nce preshow keynote speech partnership tivo call tivotogo mean people play
record program windows pcs mobile devices reflect increase trend free mult
imedia people watch want want'
```

Encoding Category column using Label encoder

In [22]:

```
en_l = LabelEncoder()
en_l.fit(finews_d['Category'])
finews_d['n_Category'] = en_l.transform(finews_d['Category'])

x = finews_d['nw_Article']
y = finews_d['n_Category']

# Perform train-test split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25, r
```

In [23]:
```
x_train.shape
```
Out[23]:
```
(1594,)
```
In [24]:
```
y_train.shape
```
Out[24]:
```
(1594,)
```
In [25]:
```
x_test.shape
```
Out[25]:
```
(532,)
```
In [26]:
```
y_test.shape
```
Out[26]:
```
(532,)
```
In [27]:
```
# Option for the user to choose between Bag of Words and TF-IDF techniques
```
In [28]:
```
method_c = input()
```
```
bow
```
In [29]:
```
if method_c == 'bow':
    bw_v = CountVectorizer()
    x_train = bw_v.fit_transform(x_train).toarray()
    x_test = bw_v.transform(x_test).toarray()
else:
    tf_id_f = TfidfVectorizer()
    x_train = tf_id_f.fit_transform(x_train).toarray()
    x_test = tf_id_f.transform(x_test).toarray()
```
In [30]:
```
x_train
```
Out[30]:
```
array([[0, 0, 0, ..., 0, 0, 0],
       [0, 2, 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)
```

Naive Bayes

In [31]:
```
nac_be = GaussianNB()
nac_be.fit(x_train, y_train)
nac_yp = nac_be.predict(x_test)
```

```
In [32]: # classification report
         print(classification_report(y_test, nac_yp))
```

```
              precision    recall  f1-score   support

           0       0.91      0.87      0.89       120
           1       0.90      0.92      0.91        87
           2       0.88      0.94      0.91        96
           3       0.96      0.96      0.96       133
           4       0.90      0.88      0.89        96

    accuracy                           0.91       532
   macro avg       0.91      0.91      0.91       532
weighted avg       0.91      0.91      0.91       532
```

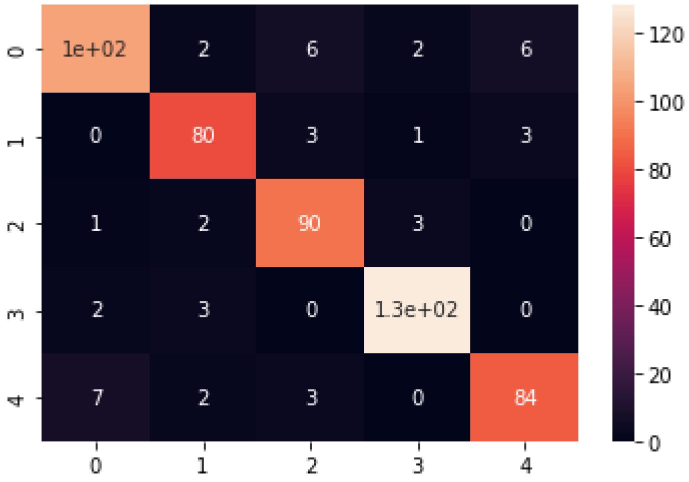
Naive bayes

- accuracy is 0.91
- precision for 0, 1, 2, 3, 4 is between 0.88 - 0.96 and so as recall

```
In [33]: # confusion matrix
         confusion_matrix(y_test, nac_yp)
```

```
Out[33]: array([[104,   2,   6,   2,   6],
                [  0,  80,   3,   1,   3],
                [  1,   2,  90,   3,   0],
                [  2,   3,   0, 128,   0],
                [  7,   2,   3,   0,  84]], dtype=int64)
```

```
In [34]: sns.heatmap(confusion_matrix(y_test, nac_yp), annot=True)
         plt.show()
```



The confusion matrix plot for naive bayes

Functionalizing code for 3 classifier models

In [35]:

```
def choose_model(model_name):
    if model_name == 'Decision Tree':
        dtc_m = DecisionTreeClassifier(random_state=9)
        dtc_m.fit(x_train, y_train)
        dtc_m_yp = dtc_m.predict(x_test)
        print(classification_report(y_test, dtc_m_yp))
        print('Confusion Matrix:')
        sns.heatmap(confusion_matrix(y_test, dtc_m_yp), annot=True)
        plt.show()
        return confusion_matrix(y_test, dtc_m_yp)

    if model_name=='Random Forest':
        rf_m = RandomForestClassifier(random_state=9)
        rf_m.fit(x_train, y_train)
        rf_m_yp = rf_m.predict(x_test)
        print(classification_report(y_test, rf_m_yp))
        print('Confusion Matrix:')
        sns.heatmap(confusion_matrix(y_test, rf_m_yp), annot=True)
        plt.show()
        return confusion_matrix(y_test, rf_m_yp)

    if model_name=='Nearest Neighbors':
        nn_m = KNeighborsClassifier(n_neighbors=4)
        nn_m.fit(x_train, y_train)
        nn_m_yp = nn_m.predict(x_test)
        print(classification_report(y_test, nn_m_yp))
        print('Confusion Matrix:')
        sns.heatmap(confusion_matrix(y_test, nn_m_yp), annot=True)
        plt.show()
        return confusion_matrix(y_test, nn_m_yp)
```

Decision Tree

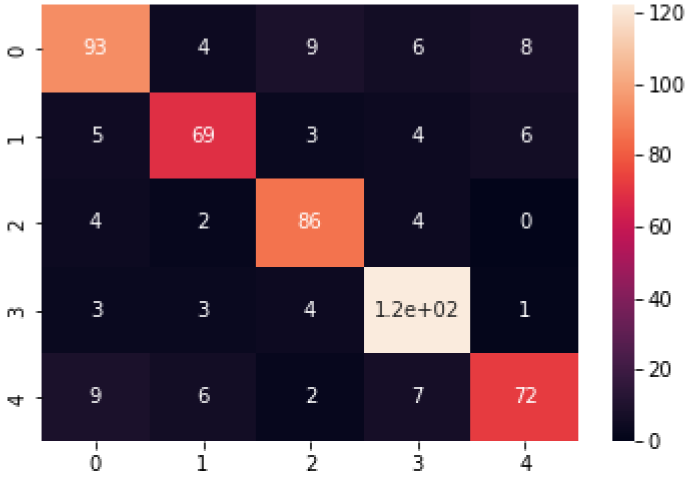
```
In [36]: choose_model('Decision Tree')
```

```
              precision    recall  f1-score   support

           0       0.82      0.78      0.79       120
           1       0.82      0.79      0.81        87
           2       0.83      0.90      0.86        96
           3       0.85      0.92      0.88       133
           4       0.83      0.75      0.79        96

    accuracy                           0.83       532
   macro avg       0.83      0.83      0.83       532
weighted avg       0.83      0.83      0.83       532

Confusion Matrix:
```



```
Out[36]: array([[ 93,   4,   9,   6,   8],
                [  5,  69,   3,   4,   6],
                [  4,   2,  86,   4,   0],
                [  3,   3,   4, 122,   1],
                [  9,   6,   2,   7,  72]], dtype=int64)
```

Decision Tree

- accuracy:0.83
- precision between 0.82-0.85
- recall between 0.75-0.92

Nearest Neighbors

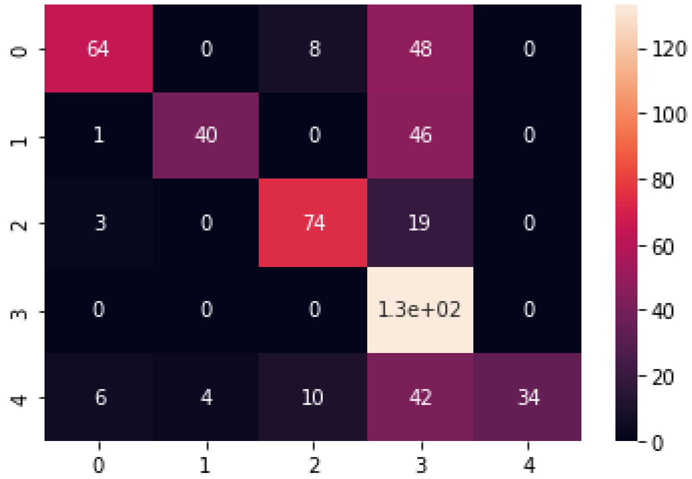
```
In [37]: choose_model('Nearest Neighbors')
```

```
              precision    recall  f1-score   support

           0       0.86      0.53      0.66       120
           1       0.91      0.46      0.61        87
           2       0.80      0.77      0.79        96
           3       0.46      1.00      0.63       133
           4       1.00      0.35      0.52        96

    accuracy                           0.65       532
   macro avg       0.81      0.62      0.64       532
weighted avg       0.78      0.65      0.64       532

Confusion Matrix:
```



```
Out[37]: array([[ 64,   0,   8,  48,   0],
               [  1,  40,   0,  46,   0],
               [  3,   0,  74,  19,   0],
               [  0,   0,   0, 133,   0],
               [  6,   4,  10,  42,  34]], dtype=int64)
```

Nearest Neighbors

- accuract is 0.65
- precision for 3 is 0.46
- recall for 4 is 0.35

Random Forest

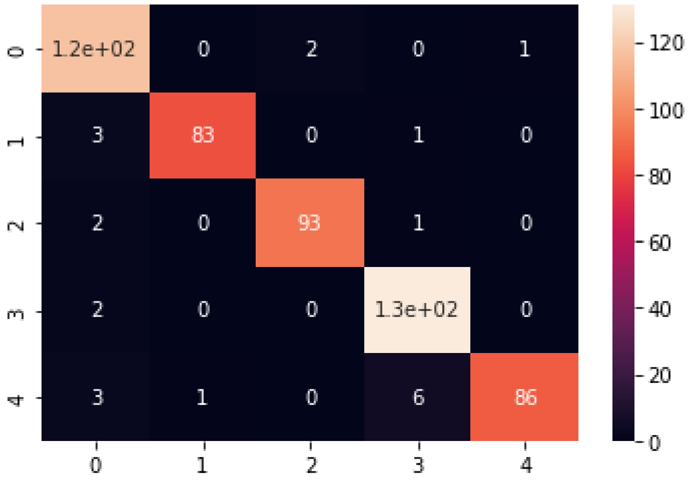
```
In [38]: choose_model('Random Forest')
```

```
              precision    recall  f1-score   support

           0       0.92      0.97      0.95       120
           1       0.99      0.95      0.97        87
           2       0.98      0.97      0.97        96
           3       0.94      0.98      0.96       133
           4       0.99      0.90      0.94        96

    accuracy                           0.96       532
   macro avg       0.96      0.96      0.96       532
weighted avg       0.96      0.96      0.96       532

Confusion Matrix:
```



```
Out[38]: array([[117,   0,   2,   0,   1],
                [  3,  83,   0,   1,   0],
                [  2,   0,  93,   1,   0],
                [  2,   0,   0, 131,   0],
                [  3,   1,   0,   6,  86]], dtype=int64)
```

Random Forest

- accuracy: 0.96
- precision for 0, 1, 2, 3, 4 is between 0.92-0.99 and recall between 0.94-0.97

Insights

- Random Forest has the highest accuracy of 0.96
- The second highest accuracy is for naive bayes which is 0.91

Questionnaire:

- How many news articles are present in the dataset we have?
  - Ans: 2225 news articles are present in the dataset

- Most of the news articles are from __ category.

  - Most of the news articles are from Sports category.

- Only __ no. of articles belong to the 'Technology' category
  - Only 401 no. of articles belong to the 'Technology' category

- What are Stop Words and why should they be removed from the text data?
  - Ans: Stop words are something that are useful while forming sentence. They need to be removed because they are not required while working on NLP projects.

- Explain the difference between Stemming and Lemmatization.
  - Ans: Stemming usually removes suffix lemmatization sees if after removing suffix does the word make sense or not

- Which of the techniques Bag of Words or TF-IDE is considered to be more efficient than the other?
  - Ans:TF-IDE

- Whats the shape of train & test data sets after performing 75:25 split.
  - Ans: shape of x_train:(1594,), y_train:(1594,), x_test:(532,), y_test:(532,)

- Which of the following found to be the best performming model a. Random Forest b. Nearest Neighbors c.Naive Bayes
  - Ans: a. Random Forest

- According to this particular use case, both precision and recall are equally important. (T/F)
  - Ans:T