Reference

- Used the following video https://www.youtube.com/watch?v=F0ys2UxRQ8I (https://www.youtube.com/watch?v=F0ys2UxRQ8I)
- Added some comments and made a few changes

Stock Sentiment Analysis using News Headlines

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

In [2]:
```
df=pd.read_csv('data/Data.csv', encoding = "ISO-8859-1")
```

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

Out[3]:

	Date	Label	Top1	Top2	Top3	Top4	Top5	Top6	Top7
0	2000-01-03	0	A 'hindrance to operations': extracts from the...	Scorecard	Hughes' instant hit buoys Blues	Jack gets his skates on at ice-cold Alex	Chaos as Maracana builds up for United	Depleted Leicester prevail as Elliott spoils E...	Hungry Spurs sense rich pickings
1	2000-01-04	0	Scorecard	The best lake scene	Leader: German sleaze inquiry	Cheerio, boyo	The main recommendations	Has Cubie killed fees?	Has Cubie killed fees?
2	2000-01-05	0	Coventry caught on counter by Flo	United's rivals on the road to Rio	Thatcher issues defence before trial by video	Police help Smith lay down the law at Everton	Tale of Trautmann bears two more retellings	England on the rack	Pakistan retaliate with call for video of Walsh
3	2000-01-06	1	Pilgrim knows how to progress	Thatcher facing ban	McIlroy calls for Irish fighting spirit	Leicester bin stadium blueprint	United braced for Mexican wave	Auntie back in fashion, even if the dress look...	Shoaib appeal goes to the top
4	2000-01-07	1	Hitches and Horlocks	Beckham off but United survive	Breast cancer screening	Alan Parker	Guardian readers: are you all whingers?	Hollywood Beyond	Ashes and diamonds

5 rows × 27 columns

In [4]:
```
train = df[df['Date'] < '20150101']
test = df[df['Date'] > '20141231']
```

In [5]:
```
# Removing punctuations
data=train.iloc[:,2:27]
data.replace("[^a-zA-Z]"," ",regex=True, inplace=True)

# Renaming column names for ease of access
list1= [i for i in range(25)]
new_Index=[str(i) for i in list1]
data.columns= new_Index
data.head(5)
```

Out[5]:

	0	1	2	3	4	5	6	7	
0	A hindrance to operations extracts from the...	Scorecard	Hughes instant hit buoys Blues	Jack gets his skates on at ice cold Alex	Chaos as Maracana builds up for United	Depleted Leicester prevail as Elliott spoils E...	Hungry Spurs sense rich pickings	Gunners so wide of an easy target	g St
1	Scorecard	The best lake scene	Leader German sleaze inquiry	Cheerio boyo	The main recommendations	Has Cubie killed fees	Has Cubie killed fees	Has Cubie killed fees	H fu F Han
2	Coventry caught on counter by Flo	United s rivals on the road to Rio	Thatcher issues defence before trial by video	Police help Smith lay down the law at Everton	Tale of Trautmann bears two more retellings	England on the rack	Pakistan retaliate with call for video of Walsh	Cullinan continues his Cape monopoly	M put out
3	Pilgrim knows how to progress	Thatcher facing ban	McIlroy calls for Irish fighting spirit	Leicester bin stadium blueprint	United braced for Mexican wave	Auntie back in fashion even if the dress look...	Shoaib appeal goes to the top	Hussain hurt by shambles but lays blame on e...	En ded dis
4	Hitches and Horlocks	Beckham off but United survive	Breast cancer screening	Alan Parker	Guardian readers are you all whingers	Hollywood Beyond	Ashes and diamonds	Whingers a formidable minority	p

5 rows × 25 columns

In [6]:
```
# Convertng headlines to lower case
for index in new_Index:
    data[index]=data[index].str.lower()
data.head(1)
```

Out[6]:

	0	1	2	3	4	5	6	7	8	9	..
0	a hindrance to operations extracts from the...	scorecard	hughes instant hit buoys blues	jack gets his skates on at ice cold alex	chaos as maracana builds up for united	depleted leicester prevail as elliott spoils e...	hungry spurs sense rich pickings	gunners so wide of an easy target	derby raise a glass to strupar s debut double	southgate strikes leeds pay the penalty	..

1 rows × 25 columns

In [7]:
```
' '.join(str(x) for x in data.iloc[1,0:25])
```

Out[7]:
```
'scorecard the best lake scene leader  german sleaze inquiry cheerio  boy
o the main recommendations has cubie killed fees  has cubie killed fees
has cubie killed fees  hopkins  furious  at foster s lack of hannibal app
etite has cubie killed fees  a tale of two tails i say what i like and i
like what i say elbows  eyes and nipples task force to assess risk of ast
eroid collision how i found myself at last on the critical list the timin
g of their lives dear doctor irish court halts ira man s extradition to n
orthern ireland burundi peace initiative fades after rebels reject mandel
a as mediator pe points the way forward to the ecb campaigners keep up pr
essure on nazi war crimes suspect jane ratcliffe yet more things you woul
dn t know without the movies millennium bug fails to bite'
```

In [8]:
```
headlines = []
for row in range(0,len(data.index)):
    headlines.append(' '.join(str(x) for x in data.iloc[row,0:25]))
```

In [9]:
```
headlines[0]
```

Out[9]:
```
'a  hindrance to operations  extracts from the leaked reports scorecard
hughes  instant hit buoys blues jack gets his skates on at ice cold alex
chaos as maracana builds up for united depleted leicester prevail as elli
ott spoils everton s party hungry spurs sense rich pickings gunners so wi
de of an easy target derby raise a glass to strupar s debut double southg
ate strikes  leeds pay the penalty hammers hand robson a youthful lesson
saints party like it s      wear wolves have turned into lambs stump mike
catches testy gough s taunt langer escapes to hit     flintoff injury pil
es on woe for england hunters threaten jospin with new battle of the somm
e kohl s successor drawn into scandal the difference between men and wome
n sara denver  nurse turned solicitor diana s landmine crusade put tories
in a panic yeltsin s resignation caught opposition flat footed russian ro
ulette sold out recovering a title'
```

In [10]:
```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.ensemble import RandomForestClassifier
```

In [11]:
```
## implement BAG OF WORDS
countvector=CountVectorizer(ngram_range=(2,2))
traindataset=countvector.fit_transform(headlines)
```

In [12]:
```
# implement RandomForest Classifier
randomclassifier=RandomForestClassifier(n_estimators=200,criterion='entropy
randomclassifier.fit(traindataset,train['Label'])
```

Out[12]:
```
RandomForestClassifier(criterion='entropy', n_estimators=200)
```

In [13]:
```
## Predict for the Test Dataset
test_transform= []
for row in range(0,len(test.index)):
    test_transform.append(' '.join(str(x) for x in test.iloc[row,2:27]))
test_dataset = countvector.transform(test_transform)
predictions = randomclassifier.predict(test_dataset)
```

In [14]:
```
## Import library to check accuracy
from sklearn.metrics import classification_report,confusion_matrix,accuracy
```

In [15]:
```
matrix=confusion_matrix(test['Label'],predictions)
print(matrix)
score=accuracy_score(test['Label'],predictions)
print(score)
report=classification_report(test['Label'],predictions)
print(report)
```

```
[[140  46]
 [  9 183]]
0.8544973544973545
              precision    recall  f1-score   support

           0       0.94      0.75      0.84       186
           1       0.80      0.95      0.87       192

    accuracy                           0.85       378
   macro avg       0.87      0.85      0.85       378
weighted avg       0.87      0.85      0.85       378
```

In [ ]: