In [25]:

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

In [26]:

```
df = pd.read_csv('../TextFiles/moviereviews.tsv', sep = '\t')
df.head()
```

Out[26]:

	label	review
0	neg	how do films like mouse hunt get into theatres...
1	neg	some talented actresses are blessed with a dem...
2	pos	this has been an extraordinary year for austra...
3	pos	according to hollywood movies made in last few...
4	neg	my first press screening of 1998 and already i...

In [27]:

```
df.isnull().sum()
```

Out[27]:

```
label      0
review    35
dtype: int64
```

In [28]:

```
# 35 of the labels are null
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 2 columns):
label     2000 non-null object
review    1965 non-null object
dtypes: object(2)
memory usage: 31.3+ KB
```

In [29]:

```
df.dropna(inplace = True)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1965 entries, 0 to 1999
Data columns (total 2 columns):
label     1965 non-null object
review    1965 non-null object
dtypes: object(2)
memory usage: 46.1+ KB
```

In [30]:

```
# Remove blanks
blanks = []
for i, lb, rv in df.itertuples():
    if type(rv) == str:
        if rv.isspace():
            blanks.append(i)

print(len(blanks))
```

```
27
```

In [31]:

```
df.drop(blanks, inplace = True)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1938 entries, 0 to 1999
Data columns (total 2 columns):
label     1938 non-null object
review    1938 non-null object
dtypes: object(2)
memory usage: 45.4+ KB
```

In [32]:

```
df['label'].value_counts()
```

Out[32]:

```
neg    969
pos    969
Name: label, dtype: int64
```

In [33]:

```
from nltk.sentiment.vader import SentimentIntensityAnalyzer

sid = SentimentIntensityAnalyzer()
```

In [34]:

```
# Use sid to append a comp_score to the dataset

sid.polarity_scores(df.iloc[0]['review'])
```

Out[34]:

```
{'neg': 0.121, 'neu': 0.778, 'pos': 0.101, 'compound': -0.9125}
```

In [35]:

```
def calc_comp_score(rev):
    adict = sid.polarity_scores(rev)
    compound = adict['compound']
    if compound >= 0:
        return('pos')
    else:
        return('neg')
```

In [36]:
```
df.iloc[0]['review']
```

Out[36]:
```
'how do films like mouse hunt get into theatres ? \r\nisn\'t there a l
aw or something ? \r\nthis diabolical load of claptrap from steven spe
ilberg\'s dreamworks studio is hollywood family fare at its deadly wor
st . \r\nmouse hunt takes the bare threads of a plot and tries to prop
it up with overacting and flat-out stupid slapstick that makes comedie
s like jingle all the way look decent by comparison . \r\nwriter adam
rifkin and director gore verbinski are the names chiefly responsible f
or this swill . \r\nthe plot , for what its worth , concerns two broth
ers ( nathan lane and an appalling lee evens ) who inherit a poorly ru
n string factory and a seemingly worthless house from their eccentric
father . \r\ndeciding to check out the long-abandoned house , they soo
n learn that it\'s worth a fortune and set about selling it in auction
to the highest bidder . \r\nbut battling them at every turn is a very
smart mouse , happy with his run-down little abode and wanting it to s
tay that way . \r\nthe story alternates between unfunny scenes of the
brothers bickering over what to do with their inheritance and endless
action sequences as the two take on their increasingly determined furr
y foe . \r\nwhatever promise the film starts with soon deteriorates in
to boring dialogue , terrible overacting , and increasingly uninspired
slapstick that becomes all sound and fury , signifying nothing . \r\nt
he script becomes so unspeakably bad that the best line poor lee evens
can utter after another run in with the rodent is : " i hate that mous
e " . \r\noh cringe ! \r\nthis is home alone all over again , and ten
times worse . \r\none touching scene early on is worth mentioning . \r
\nwe follow the mouse through a maze of walls and pipes until he arriv
es at his makeshift abode somewhere in a wall . \r\nhe jumps into a ti
ny bed , pulls up a makeshift sheet and snuggles up to sleep , seeming
ly happy and just wanting to be left alone . \r\nit\'s a magical littl
e moment in an otherwise soulless film . \r\na message to speilberg :
if you want dreamworks to be associated with some kind of artistic cre
dibility , then either give all concerned in mouse hunt a swift kick u
p the arse or hire yourself some decent writers and directors . \r\nth
is kind of rubbish will just not do at all . \r\n'
```

In [37]:
```
print(calc_comp_score(df.iloc[0]['review']))
```

```
neg
```

In [38]:
```
# Create a new column called comp_score

df['comp_score'] = df['review'].apply(calc_comp_score)
```

In [39]:
```
df.head()
```

Out[39]:

	label	review	comp_score
0	neg	how do films like mouse hunt get into theatres...	neg
1	neg	some talented actresses are blessed with a dem...	neg
2	pos	this has been an extraordinary year for austra...	pos
3	pos	according to hollywood movies made in last few...	pos
4	neg	my first press screening of 1998 and already i...	neg

Perform a comparison between original label and comp_score

In [40]:
```
from sklearn.metrics import classification_report, confusion_matrix
from sklearn.metrics import accuracy_score

print(classification_report(df['label'], df['comp_score']))
```

```
              precision    recall  f1-score   support

         neg       0.72      0.44      0.55       969
         pos       0.60      0.83      0.70       969

   micro avg       0.64      0.64      0.64      1938
   macro avg       0.66      0.64      0.62      1938
weighted avg       0.66      0.64      0.62      1938
```

In [41]:
```
print(confusion_matrix(df['label'], df['comp_score']))
```

```
[[427 542]
 [162 807]]
```

In [42]:
```
print(accuracy_score(df['label'], df['comp_score']))
```

```
0.6367389060887513
```

Conclusion: It looks like Vader could not judge the movie reviews accurately. Understanding human semantics is the biggest challenge in Sentiment Analysis. Many of the reviews look misleading, i.e. they have positive things to say at the beginning and last sentence has negative review. This is confusing to Vader.

In [ ]: