SentimentAnalysis

June 11, 2022

1 Sentiment Analysis

The Sentiment analysis (also known as opinion mining or emotion AI) is the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information (from Wikipedia).

In this example, we have to predict the number of positive and negative reviews based on sentiments. For sake of semplicity we'll use just one model based on the Naive Bayes approach. This to text classification is a good choice when the amount of training data is limited.

1.0.1 Import data set

[ ]:
```
import pandas as pd

# from: https://www.kaggle.com/datasets/lakshmi25npathi/
 ↪imdb-dataset-of-50k-movie-reviews?select=IMDB+Dataset.csv
dfMovies = pd.read_csv(r'./data/IMDB Dataset.csv', engine="c", encoding='utf-8')
```

1.0.2 Basic EDA and cleaning data

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

[ ]:

	review	sentiment
0	One of the other reviewers has mentioned that …	positive
1	A wonderful little production. The…	positive
2	I thought this was a wonderful way to spend ti…	positive
3	Basically there's a family where a little boy …	negative
4	Petter Mattei's "Love in the Time of Money" is…	positive

We note that some review beed to be cleaned (for example see the second review).

[ ]:
```
# Just use python variable replacement syntax to make the text dynamic.
from IPython.display import Markdown as md

md(f"The IMDb data set consists of {dfMovies.shape[1]} different parameters of␣
 ↪wine which was measured for {dfMovies.shape[0]} review samples.")
```

[ ]:
The IMDb data set consists of 2 different parameters of wine which was measured for 50000 review samples.

Type data and memory usage

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50000 entries, 0 to 49999
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   review     50000 non-null  object
 1   sentiment  50000 non-null  object
dtypes: object(2)
memory usage: 781.4+ KB
```

[ ]:
```
print (dfMovies.applymap(type))
```

```
                review      sentiment
0      <class 'str'>  <class 'str'>
1      <class 'str'>  <class 'str'>
2      <class 'str'>  <class 'str'>
3      <class 'str'>  <class 'str'>
4      <class 'str'>  <class 'str'>
…                …              …
49995  <class 'str'>  <class 'str'>
49996  <class 'str'>  <class 'str'>
49997  <class 'str'>  <class 'str'>
49998  <class 'str'>  <class 'str'>
49999  <class 'str'>  <class 'str'>

[50000 rows x 2 columns]
```

All data are string even if in memory the are considered to be object data. The next step is searching for missing, NA and null values. First of all verify is there are null or empty values.

[ ]:
```
(dfMovies.isnull() | dfMovies.empty).sum()
```

[ ]:
```
review       0
sentiment    0
dtype: int64
```

Second we can remove potential unwanted characters like HTML tags and special characters

[ ]:
```
import regex
```

```
def cleanText(text):
    # HTML tags
    text = regex.sub(r"<[^<]+?>","",text)
    # Special chars
    text = regex.sub(r"[^a-zA-Z0-9 ]","",text)
    text = text.lower()
    return text
```

[ ]:
```
dfMovies["review"] = dfMovies["review"].apply(cleanText)
```

As we can see, (for example) the second revies are cleaned now.

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

[ ]:

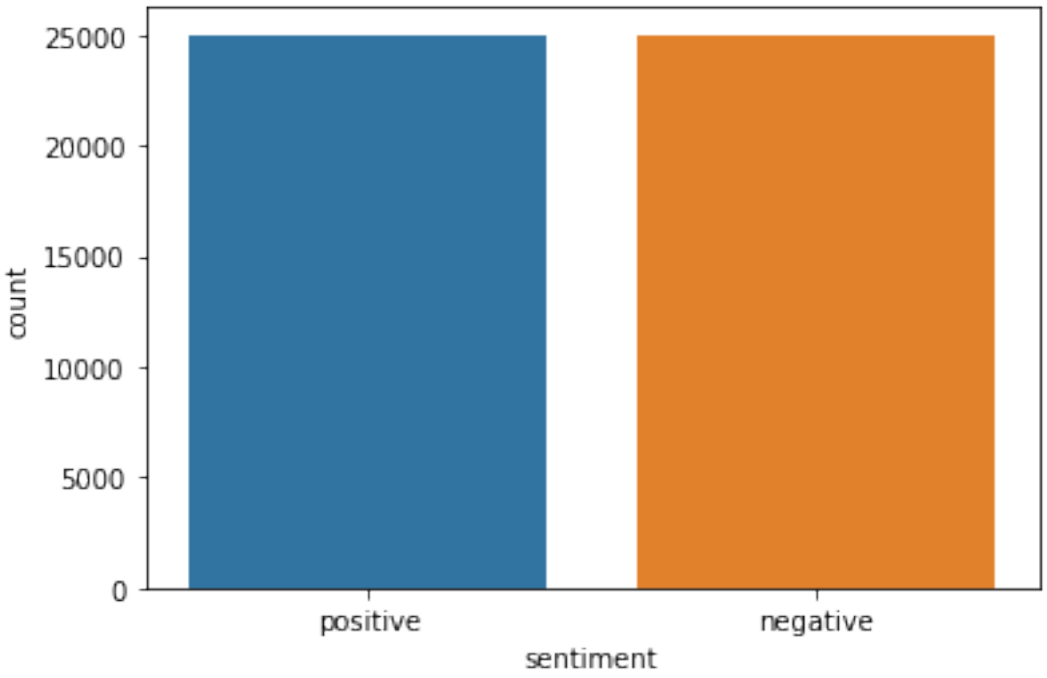
	review	sentiment
0	one of the other reviewers has mentioned that …	positive
1	a wonderful little production the filming tech…	positive
2	i thought this was a wonderful way to spend ti…	positive
3	basically theres a family where a little boy j…	negative
4	petter matteis love in the time of money is a …	positive

Let's show the summary statistics.

[ ]:
```
import seaborn as sns

sns.countplot(x=dfMovies['sentiment'])
```

[ ]: <AxesSubplot:xlabel='sentiment', ylabel='count'>



It seems to be an ex-equo in both target values: the dataset is balanced. Now we try make a visual representation of text data in order to show the importance of each word by font size or color.

[ ]:
```
from wordcloud import WordCloud
def makeWordCloud(bkColor, w, h, series, title, ax):
    wcl = WordCloud(background_color=bkColor,
                    width=w,
                    height=h).generate(" ".join(series))
    ax.imshow(wcl)
    ax.axis('off')
    ax.set_title(title, fontsize=40)
```

[ ]:
```
import matplotlib.pyplot as plt

positive = dfMovies[dfMovies['sentiment']=="positive"]['review']
negative = dfMovies[dfMovies['sentiment']=="negative"]['review']

fig, (ax1, ax2) = plt.subplots(1, 2, figsize=[26, 8])
makeWordCloud("white", 600, 400, positive, "Positive", ax1)
makeWordCloud("white", 600, 400, negative, "Negative", ax2)
plt.show()
```

We note the most used words are "movie" and "film"; maybe can get rid off of these words without any consequences.

[ ]:
```
dfMovies["review"] = dfMovies["review"].str.replace("movie", "")
dfMovies["review"] = dfMovies["review"].str.replace("film", "")
```

A quick check:

[ ]:
```
dfMovies['review'].str.contains('movie').eq(True).all()
```

[ ]: False

[ ]:
```
dfMovies['review'].str.contains('film').eq(True).all()
```

[ ]: False

1.0.3 Train and test the model

Computers cannot understand words, they can only understand numbers. For this purpose, Countvectorizer converts text to numerical data (more specifically, the text is transformed to a sparse matrix) by replacing all-knowing words with the number of times they appear. In a nutshell, the model will count how many times the word 'good' appears in a positive sentence, and then divide this number by how many times this word appears at all. For the negative score, "1 - Positive Score". This process is called vectorization. Let's start with split review from sentiment.

[ ]:
```
X = dfMovies['review']
y = dfMovies['sentiment']
```

In the following code, you can see: CountVectorizer(ngram_range=(1,2)) This statement converts the text to lowercase (by default). About the ngram_range parameter, it lets you decide the sequence length of consecutive words in the given text. In the example, it will pick the unigram (that is only a single word) and bigram (group of two consecutive words).

[ ]:
```
from sklearn.feature_extraction.text import CountVectorizer

vect = CountVectorizer(ngram_range=(1,2))
```

For the sake of clarity, fit and transform statements are keep separated. In the following line we show the mapping of terms.

[ ]:
```
vect.fit(X).vocabulary_
```

[ ]:
```
{'one': 1734328,
 'of': 1683535,
 'the': 2381259,
 'other': 1765922,
 'reviewers': 2012610,
 'has': 1068945,
 'mentioned': 1527800,
 'that': 2367502,
 'after': 71391,
 'watching': 2670669,
 'just': 1328151,
 'oz': 1791785,
```

```
'episode': 786389,
'youll': 2798030,
'be': 274166,
'hooked': 1151869,
'they': 2446303,
'are': 189381,
'right': 2020486,
'as': 207474,
'this': 2456960,
'is': 1263189,
'exactly': 813914,
'what': 2697627,
'happened': 1061251,
'with': 2739942,
'methe': 1533210,
'first': 889500,
'thing': 2451188,
'struck': 2287835,
'me': 1512146,
'about': 31528,
'was': 2655872,
'its': 1292308,
'brutality': 384934,
'and': 126652,
'unflinching': 2582098,
'scenes': 2076091,
'violence': 2635761,
'which': 2709160,
'set': 2123626,
'in': 1202074,
'from': 944023,
'word': 2764654,
'go': 1001363,
'trust': 2551966,
'not': 1657480,
'show': 2152248,
'for': 909902,
'faint': 842876,
'hearted': 1091320,
'or': 1750884,
'timid': 2491723,
'pulls': 1928654,
'no': 1644009,
'punches': 1929433,
'regards': 1983841,
'to': 2495467,
'drugs': 724130,
```

```
'sex': 2127656,
'hardcore': 1064327,
'classic': 497496,
'use': 2604838,
'wordit': 2765355,
'called': 418072,
'nickname': 1639431,
'given': 995975,
'oswald': 1765774,
'maximum': 1507027,
'security': 2097896,
'state': 2258908,
'penitentary': 1820574,
'it': 1282512,
'focuses': 904900,
'mainly': 1471009,
'on': 1725549,
'emerald': 761824,
'city': 493926,
'an': 120335,
'experimental': 827137,
'section': 2097580,
'prison': 1905075,
'where': 2706165,
'all': 90501,
'cells': 452276,
'have': 1075645,
'glass': 998597,
'fronts': 951566,
'face': 838135,
'inwards': 1260100,
'so': 2197901,
'privacy': 1905830,
'high': 1116814,
'agenda': 78431,
'em': 760478,
'home': 1147567,
'manyaryans': 1490543,
'muslims': 1601957,
'gangstas': 967208,
'latinos': 1378784,
'christians': 486586,
'italians': 1290142,
'irish': 1261037,
'moreso': 1575364,
'scuffles': 2091759,
'death': 626054,
```

'stares': 2254480,
'dodgy': 696145,
'dealings': 625404,
'shady': 2131259,
'agreements': 80972,
'never': 1630277,
'far': 853375,
'awayi': 246370,
'would': 2774995,
'say': 2068097,
'main': 1470178,
'appeal': 183591,
'due': 727288,
'fact': 839881,
'goes': 1004562,
'shows': 2156004,
'wouldnt': 2776736,
'dare': 613187,
'forget': 923508,
'pretty': 1899279,
'pictures': 1844063,
'painted': 1795970,
'mainstream': 1471335,
'audiences': 238109,
'charm': 470166,
'romanceoz': 2035406,
'doesnt': 697845,
'mess': 1530930,
'around': 200939,
'ever': 804851,
'saw': 2066868,
'nasty': 1616499,
'surreal': 2320956,
'couldnt': 569798,
'ready': 1961992,
'but': 396794,
'watched': 2669669,
'more': 1570178,
'developed': 655752,
'taste': 2343659,
'got': 1016424,
'accustomed': 45022,
'levels': 1401167,
'graphic': 1023359,
'injustice': 1234394,
'crooked': 593097,
'guards': 1039151,

```
'wholl': 2722819,
'sold': 2206325,
'out': 1773945,
'nickel': 1639324,
'inmates': 1234646,
'kill': 1346627,
'order': 1759608,
'get': 982350,
'away': 245575,
'well': 2685480,
'mannered': 1486056,
'middle': 1537107,
'class': 496795,
'being': 302105,
'turned': 2556316,
'into': 1250831,
'bitches': 332227,
'their': 2425480,
'lack': 1365007,
'street': 2283214,
'skills': 2180333,
'experience': 825850,
'you': 2794556,
'may': 1507169,
'become': 289143,
'comfortable': 524242,
'uncomfortable': 2573666,
'viewingthats': 2632737,
'if': 1185864,
'can': 424132,
'touch': 2527458,
'your': 2800682,
'darker': 614530,
'side': 2160237,
'one of': 1736817,
'of the': 1708018,
'the other': 2407758,
'other reviewers': 1768658,
'reviewers has': 2012681,
'has mentioned': 1071100,
'mentioned that': 1528045,
'that after': 2367787,
'after watching': 73677,
'watching just': 2671424,
'just oz': 1330784,
'oz episode': 1791817,
'episode youll': 787065,
```

```
'youll be': 2798051,
'be hooked': 277458,
'hooked they': 1151940,
'they are': 2446419,
'are right': 194642,
'right as': 2020561,
'as this': 218555,
'this is': 2461735,
'is exactly': 1267661,
'exactly what': 814349,
'what happened': 2698766,
'happened with': 1061536,
'with methe': 2747268,
'methe first': 1533238,
'first thing': 891727,
'thing that': 2452056,
'that struck': 2377685,
'struck me': 2287867,
'me about': 1512189,
'about oz': 35043,
'oz was': 1791891,
'was its': 2660027,
'its brutality': 1293092,
'brutality and': 384939,
'and unflinching': 155899,
'unflinching scenes': 2582118,
'scenes of': 2076806,
'of violence': 1709949,
'violence which': 2636220,
'which set': 2711518,
'set in': 2123860,
'in right': 1212794,
'right from': 2020856,
'from the': 950259,
'the word': 2423192,
'word go': 2764885,
'go trust': 1002164,
'trust me': 2552044,
'me this': 1514093,
'is not': 1272231,
'not show': 1662120,
'show for': 2152846,
'for the': 918956,
'the faint': 2394918,
'faint hearted': 842886,
'hearted or': 1091380,
'or timid': 1758104,
```

'timid this': 2491753,
'this show': 2465207,
'show pulls': 2153441,
'pulls no': 1928692,
'no punches': 1646931,
'punches with': 1929492,
'with regards': 2749277,
'regards to': 1983876,
'to drugs': 2499863,
'drugs sex': 724330,
'sex or': 2128004,
'or violence': 1758541,
'violence its': 2635977,
'its is': 1295498,
'is hardcore': 1269181,
'hardcore in': 1064391,
'in the': 1214912,
'the classic': 2388668,
'classic use': 498604,
'use of': 2605238,
'the wordit': 2423199,
'wordit is': 2765357,
'is called': 1265102,
'called oz': 418987,
'oz as': 1791793,
'as that': 218484,
'that is': 2373062,
'is the': 1276567,
'the nickname': 2406486,
'nickname given': 1639440,
'given to': 996579,
'to the': 2509854,
'the oswald': 2407755,
'oswald maximum': 1765786,
'maximum security': 1507073,
'security state': 2098044,
'state penitentary': 2259095,
'penitentary it': 1820575,
'it focuses': 1284920,
'focuses mainly': 904919,
'mainly on': 1471177,
'on emerald': 1727623,
'emerald city': 761825,
'city an': 493965,
'an experimental': 122647,
'experimental section': 827228,
'section of': 2097640,

```
'the prison': 2410452,
'prison where': 1905391,
'where all': 2706212,
'all the': 94431,
'the cells': 2387792,
'cells have': 452299,
'have glass': 1077375,
'glass fronts': 998649,
'fronts and': 951568,
'and face': 136361,
'face inwards': 838421,
'inwards so': 1260101,
'so privacy': 2200997,
'privacy is': 1905836,
'not high': 1659942,
'high on': 1117204,
'on the': 1732117,
'the agenda': 2382732,
'agenda em': 78449,
'em city': 760494,
'city is': 494258,
'is home': 1269498,
'home to': 1148192,
'to manyaryans': 2504280,
'manyaryans muslims': 1490544,
'muslims gangstas': 1601975,
'gangstas latinos': 967212,
'latinos christians': 1378790,
'christians italians': 486633,
'italians irish': 1290168,
'irish and': 1261048,
'and moreso': 144734,
'moreso scuffles': 1575375,
'scuffles death': 2091760,
'death stares': 626656,
'stares dodgy': 2254494,
'dodgy dealings': 696166,
'dealings and': 625406,
'and shady': 151404,
'shady agreements': 2131261,
'agreements are': 80973,
'are never': 193567,
'never far': 1630881,
'far awayi': 853419,
'awayi would': 246379,
'would say': 2776178,
'say the': 2069293,
```

```
'the main': 2403612,
'main appeal': 1470208,
'appeal of': 183684,
'the show': 2414982,
'show is': 2153042,
'is due': 1267207,
'due to': 727368,
'the fact': 2394868,
'fact that': 840574,
'that it': 2373107,
'it goes': 1285139,
'goes where': 1005035,
'where other': 2707651,
'other shows': 1768879,
'shows wouldnt': 2156978,
'wouldnt dare': 2776804,
'dare forget': 613217,
'forget pretty': 923698,
'pretty pictures': 1900156,
'pictures painted': 1844233,
'painted for': 1796003,
'for mainstream': 915455,
'mainstream audiences': 1471357,
'audiences forget': 238215,
'forget charm': 923555,
'charm forget': 470225,
'forget romanceoz': 923714,
'romanceoz doesnt': 2035407,
'doesnt mess': 698352,
'mess around': 1530953,
'around the': 202298,
'the first': 2395603,
'first episode': 890213,
'episode ever': 786588,
'ever saw': 806055,
'saw struck': 2067688,
'me as': 1512307,
'as so': 217649,
'so nasty': 2200575,
'nasty it': 1616677,
'it was': 1289082,
'was surreal': 2663519,
'surreal couldnt': 2321001,
'couldnt say': 570080,
'say was': 2069397,
'was ready': 2662086,
'ready for': 1962014,
```

```
'for it': 914577,
'it but': 1283451,
'but as': 397136,
'as watched': 219499,
'watched more': 2670063,
'more developed': 1571280,
'developed taste': 655958,
'taste for': 2343722,
'for oz': 916523,
'oz and': 1791791,
'and got': 138326,
'got accustomed': 1016456,
'accustomed to': 45026,
'the high': 2398938,
'high levels': 1117138,
'levels of': 1401253,
'of graphic': 1694024,
'graphic violence': 1023526,
'violence not': 2636045,
'not just': 1660326,
'just violence': 1332299,
'violence but': 2635829,
'but injustice': 399483,
'injustice crooked': 1234405,
'crooked guards': 593127,
'guards wholl': 1039235,
'wholl be': 2722822,
'be sold': 280374,
'sold out': 2206377,
'out for': 1774788,
'for nickel': 916196,
'nickel inmates': 1639327,
'inmates wholl': 1234710,
'wholl kill': 2722832,
'kill on': 1346998,
'on order': 1730134,
'order and': 1759623,
'and get': 137962,
'get away': 982504,
'away with': 246259,
'with it': 2745788,
'it well': 1289180,
'well mannered': 2686512,
'mannered middle': 1486077,
'middle class': 1537142,
'class inmates': 497038,
'inmates being': 1234654,
```

```
'being turned': 305358,
'turned into': 2556431,
'into prison': 1253384,
'prison bitches': 1905106,
'bitches due': 332233,
'to their': 2509866,
'their lack': 2428383,
'lack of': 1365091,
'of street': 1706992,
'street skills': 2283595,
'skills or': 2180435,
'or prison': 1756372,
'prison experience': 1905168,
'experience watching': 826331,
'watching oz': 2671658,
'oz you': 1791907,
'you may': 2796247,
'may become': 1507250,
'become comfortable': 289318,
'comfortable with': 524342,
'with what': 2752526,
'what is': 2698974,
'is uncomfortable': 1277190,
'uncomfortable viewingthats': 2573773,
'viewingthats if': 2632738,
'if you': 1187890,
'you can': 2794947,
'can get': 424804,
'get in': 983352,
'in touch': 1215239,
'touch with': 2527660,
'with your': 2752802,
'your darker': 2801210,
'darker side': 614603,
'wonderful': 2760571,
'little': 1425868,
'production': 1911365,
'ing': 1231911,
'technique': 2348756,
'very': 2622180,
'unassuming': 2571589,
'oldtimebbc': 1724225,
'fashion': 856262,
'gives': 996720,
'comforting': 524391,
'sometimes': 2218016,
'discomforting': 681135,
```

```
'sense': 2114062,
'realism': 1965117,
'entire': 782976,
'piece': 1844720,
'actors': 53437,
'extremely': 834350,
'chosen': 485301,
'michael': 1535099,
'sheen': 2138765,
'only': 1741447,
'polari': 1870536,
'he': 1083522,
'voices': 2642122,
'down': 710127,
'pat': 1811945,
'too': 2518313,
'truly': 2550508,
'see': 2098515,
'seamless': 2093051,
'editing': 745742,
'guided': 1040718,
'by': 404945,
'references': 1981996,
'williams': 2733213,
'diary': 660999,
'entries': 784869,
'worth': 2773723,
'terrificly': 2359724,
'written': 2781848,
'performed': 1828390,
'masterful': 1500882,
'great': 1025556,
'masters': 1501701,
'comedy': 521895,
'his': 1127866,
'life': 1404120,
'really': 1967312,
'comes': 523670,
'things': 2452426,
'fantasy': 852773,
'guard': 1038860,
'rather': 1955438,
'than': 2362725,
'traditional': 2533113,
'dream': 718147,
'techniques': 2348888,
'remains': 1992075,
```

```
'solid': 2207469,
'then': 2436177,
'disappears': 678317,
'plays': 1859132,
'our': 1771956,
'knowledge': 1357713,
'senses': 2114820,
'particularly': 1806489,
'concerning': 539566,
'orton': 1764534,
'halliwell': 1055161,
'sets': 2124377,
'flat': 896629,
'halliwells': 1055170,
'murals': 1596534,
'decorating': 631641,
'every': 807259,
'surface': 2318546,
'terribly': 2358981,
'done': 703315,
'wonderful little': 2761147,
'little production': 1428023,
'production the': 1911938,
'the ing': 2400492,
'ing technique': 1232136,
'technique is': 2348816,
'is very': 1277656,
'very unassuming': 2625151,
'unassuming very': 2571605,
'very oldtimebbc': 2624184,
'oldtimebbc fashion': 1724226,
'fashion and': 856273,
'and gives': 138068,
'gives comforting': 996810,
'comforting and': 524393,
'and sometimes': 152444,
'sometimes discomforting': 2218197,
'discomforting sense': 681145,
'sense of': 2114324,
'of realism': 1703437,
'realism to': 1965263,
'the entire': 2394013,
'entire piece': 783591,
'piece the': 1844976,
'the actors': 2382433,
'actors are': 53523,
'are extremely': 191579,
```

```
'extremely well': 835060,
'well chosen': 2685766,
'chosen michael': 485369,
'michael sheen': 1535658,
'sheen not': 2138821,
'not only': 1661061,
'only has': 1743014,
'has got': 1070428,
'got all': 1016477,
'the polari': 2409685,
'polari but': 1870537,
'but he': 399200,
'he has': 1084966,
'has all': 1069067,
'the voices': 2421908,
'voices down': 2642162,
'down pat': 710805,
'pat too': 1812058,
'too you': 2520714,
'can truly': 425764,
'truly see': 2551414,
'see the': 2101088,
'the seamless': 2414038,
'seamless editing': 2093056,
'editing guided': 745881,
'guided by': 1040720,
'by the': 412916,
'the references': 2411884,
'references to': 1982094,
'to williams': 2511267,
'williams diary': 2733275,
'diary entries': 661006,
'entries not': 784903,
'only is': 1743243,
'is it': 1270265,
'well worth': 2687396,
'worth the': 2774220,
'the watching': 2422259,
'watching but': 2670872,
'but it': 399569,
'it is': 1285737,
'is terrificly': 1276504,
'terrificly written': 2359728,
'written and': 2781892,
'and performed': 147037,
'performed piece': 1828471,
'piece masterful': 1844886,
```

```
'masterful production': 1500948,
'production about': 1911371,
'about one': 34998,
'the great': 2397802,
'great masters': 1027444,
'masters of': 1501774,
'of comedy': 1688661,
'comedy and': 521955,
'and his': 139542,
'his life': 1132927,
'life the': 1405603,
'the realism': 2411649,
'realism really': 1965232,
'really comes': 1967807,
'comes home': 523819,
'home with': 1148272,
'with the': 2751427,
'the little': 2402979,
'little things': 1428683,
'things the': 2453209,
'the fantasy': 2395062,
'fantasy of': 853007,
'the guard': 2398034,
'guard which': 1039008,
'which rather': 2711283,
'rather than': 1956683,
'than use': 2366677,
'use the': 2605415,
'the traditional': 2419795,
'traditional dream': 2533189,
'dream techniques': 718400,
'techniques remains': 2348957,
'remains solid': 1992325,
'solid then': 2207838,
'then disappears': 2436865,
'disappears it': 678351,
'it plays': 1286995,
'plays on': 1860095,
'on our': 1730158,
'our knowledge': 1772817,
'knowledge and': 1357720,
'and our': 146308,
'our senses': 1773334,
'senses particularly': 2114877,
'particularly with': 1807182,
'the scenes': 2413637,
'scenes concerning': 2076303,
```

```
'concerning orton': 539640,
'orton and': 1764537,
'and halliwell': 138878,
'halliwell and': 1055162,
'and the': 154441,
'the sets': 2414515,
'sets particularly': 2124575,
'particularly of': 1806942,
'of their': 1708040,
'their flat': 2427476,
'flat with': 896939,
'with halliwells': 2744842,
'halliwells murals': 1055171,
'murals decorating': 1596535,
'decorating every': 631642,
'every surface': 808532,
'surface are': 2318557,
'are terribly': 195663,
'terribly well': 2359252,
'well done': 2685937,
'thought': 2473368,
'way': 2674507,
'spend': 2237129,
'time': 2485820,
'hot': 1160407,
'summer': 2310057,
'weekend': 2682991,
'sitting': 2176980,
'air': 82772,
'conditioned': 541366,
'theater': 2423901,
'lighthearted': 1408199,
'plot': 1862729,
'simplistic': 2167851,
'dialogue': 659512,
'witty': 2755922,
'characters': 465608,
'likable': 1409118,
'even': 799209,
'bread': 370556,
'suspected': 2323235,
'serial': 2119132,
'killer': 1347787,
'while': 2712230,
'some': 2208827,
'disappointed': 678503,
'when': 2702218,
```

'realize': 1966717,
'match': 1502259,
'point': 1867599,
'risk': 2024292,
'addiction': 61549,
'proof': 1917759,
'woody': 2764263,
'allen': 95703,
'still': 2267566,
'fully': 954699,
'control': 555513,
'style': 2294658,
'many': 1487865,
'us': 2602881,
'grown': 1037472,
'lovethis': 1453214,
'most': 1578045,
'id': 1182034,
'laughed': 1380331,
'at': 226784,
'woodys': 2764341,
'comedies': 521617,
'years': 2788720,
'decade': 628340,
'ive': 1301245,
'been': 292728,
'impressed': 1200138,
'scarlet': 2072836,
'johanson': 1317569,
'she': 2136167,
'managed': 1482949,
'tone': 2517079,
'her': 1100197,
'sexy': 2129718,
'image': 1192254,
'jumped': 1326692,
'average': 243002,
'spirited': 2239796,
'young': 2798292,
'womanthis': 2758962,
'crown': 594813,
'jewel': 1313084,
'career': 433981,
'wittier': 2755881,
'devil': 657145,
'wears': 2681156,
'prada': 1889464,

```
'interesting': 1246260,
'superman': 2314097,
'friends': 941546,
'thought this': 2474090,
'this was': 2467070,
'was wonderful': 2664661,
'wonderful way': 2761610,
'way to': 2676012,
'to spend': 2508902,
'spend time': 2237284,
'time on': 2487445,
'on too': 1732276,
'too hot': 2519279,
'hot summer': 1160775,
'summer weekend': 2310306,
'weekend sitting': 2683115,
'sitting in': 2177026,
'the air': 2382800,
'air conditioned': 82840,
'conditioned theater': 541381,
'theater and': 2423925,
'and watching': 157047,
'watching lighthearted': 2671476,
'lighthearted comedy': 1408224,
'comedy the': 522937,
'the plot': 2409479,
'plot is': 1863436,
'is simplistic': 1275201,
'simplistic but': 2167867,
'but the': 402098,
'the dialogue': 2391781,
'dialogue is': 659808,
'is witty': 1278155,
'witty and': 2755928,
'the characters': 2388016,
'characters are': 465730,
'are likable': 193034,
'likable even': 1409184,
'even the': 802637,
'the well': 2422518,
'well bread': 2685687,
'bread suspected': 370611,
'suspected serial': 2323271,
'serial killer': 2119176,
'killer while': 1348545,
'while some': 2714111,
'some may': 2211655,
```

```
'may be': 1507246,
'be disappointed': 275985,
'disappointed when': 678740,
'when they': 2705539,
'they realize': 2448182,
'realize this': 1966829,
'not match': 1660677,
'match point': 1502461,
'point risk': 1868144,
'risk addiction': 2024296,
'addiction thought': 61608,
'thought it': 2473712,
'was proof': 2661894,
'proof that': 1917815,
'that woody': 2379240,
'woody allen': 2764266,
'allen is': 95814,
'is still': 1275858,
'still fully': 2268314,
'fully in': 954893,
'in control': 1205317,
'control of': 555678,
'the style': 2417540,
'style many': 2295032,
'many of': 1489462,
'of us': 1709582,
'us have': 2603509,
'have grown': 1077448,
'grown to': 1037559,
'to lovethis': 2504049,
'lovethis was': 1453222,
'was the': 2663723,
'the most': 2405523,
'most id': 1579330,
'id laughed': 1182215,
'laughed at': 1380343,
'at one': 229583,
'of woodys': 1710812,
'woodys comedies': 2764346,
'comedies in': 521709,
'in years': 1216338,
'years dare': 2788975,
'dare say': 613244,
'say decade': 2068407,
'decade while': 628483,
'while ive': 2713256,
'ive never': 1301409,
```

```
'never been': 1630431,
'been impressed': 294293,
'impressed with': 1200229,
'with scarlet': 2749820,
'scarlet johanson': 2072856,
'johanson in': 1317571,
'in this': 1215016,
'this she': 2465139,
'she managed': 2137325,
'managed to': 1482989,
'to tone': 2510128,
'tone down': 2517137,
'down her': 710538,
'her sexy': 1105394,
'sexy image': 2129903,
'image and': 1192263,
'and jumped': 141625,
'jumped right': 1326725,
'right into': 2020961,
'into average': 1251039,
'average but': 243077,
'but spirited': 401762,
'spirited young': 2239871,
'young womanthis': 2799994,
'womanthis may': 2758964,
'may not': 1507562,
'not be': 1657957,
'be the': 280927,
'the crown': 2390519,
'crown jewel': 594835,
'jewel of': 1313114,
'of his': 1695029,
'his career': 1129229,
'career but': 434050,
'was wittier': 2664647,
'wittier than': 2755887,
'than devil': 2363755,
'devil wears': 657314,
'wears prada': 2681277,
'prada and': 1889466,
'and more': 144713,
'more interesting': 1572446,
'interesting than': 1247280,
'than superman': 2366292,
'superman great': 2314154,
'great comedy': 1026179,
'comedy to': 522967,
```

```
'to go': 2501532,
'go see': 1002019,
'see with': 2101381,
'with friends': 2744257,
'basically': 269302,
'theres': 2442164,
'family': 847553,
'boy': 364485,
'jake': 1304566,
'thinks': 2455743,
'zombie': 2809339,
'closet': 507440,
'parents': 1801830,
'fighting': 878946,
'timethis': 2491476,
'slower': 2188973,
'soap': 2202555,
'opera': 1747815,
'suddenly': 2305964,
'decides': 630618,
'rambo': 1949536,
'zombieok': 2809896,
'youre': 2803170,
'going': 1005312,
'make': 1472939,
'must': 1602078,
'decide': 630341,
'thriller': 2477444,
'drama': 714930,
'watchable': 2669331,
'divorcing': 691429,
'arguing': 198242,
'like': 1409410,
'real': 1962220,
'we': 2677864,
'totally': 2526095,
'ruins': 2045007,
'expected': 824546,
'boogeyman': 352288,
'similar': 2165487,
'instead': 1240308,
'meaningless': 1516092,
'spots3': 2244852,
'10': 343,
'playing': 1857920,
'descent': 647623,
'dialogs': 659373,
```

```
'shots': 2150176,
'ignore': 1188340,
'them': 2431507,
'basically theres': 269932,
'theres family': 2442432,
'family where': 848766,
'where little': 2707341,
'little boy': 1426177,
'boy jake': 364834,
'jake thinks': 1304667,
'thinks theres': 2455901,
'theres zombie': 2443192,
'zombie in': 2809558,
'in his': 1208267,
'his closet': 1129547,
'closet his': 507473,
'his parents': 1134217,
'parents are': 1801861,
'are fighting': 191687,
'fighting all': 878959,
'the timethis': 2419336,
'timethis is': 2491481,
'is slower': 1275344,
'slower than': 2189002,
'than soap': 2366105,
'soap opera': 2202600,
'opera and': 1747833,
'and suddenly': 153535,
'suddenly jake': 2306189,
'jake decides': 1304591,
'decides to': 630671,
'to become': 2496760,
'become rambo': 289977,
'rambo and': 1949538,
'and kill': 141900,
'kill the': 1347110,
'the zombieok': 2423821,
'zombieok first': 2809897,
'first of': 890956,
'of all': 1684733,
'all when': 94888,
'when youre': 2705930,
'youre going': 2803477,
'going to': 1005735,
'to make': 2504181,
'make you': 1474719,
'you must': 2796339,
```

```
 'must decide': 1602225,
 …}
```

and then, it will transform the input data.

[ ]:
```
X = vect.transform(X)
#print(X)
```

Now we can split in train and test datasets, in order to use validation data over test data.

[ ]:
```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y)
```

We'll use the BernoulliNB (NB stands for Naive Bayes) will generate our classification model.

The Naive Bayes classifier uses the Bayes Theorem that represents the probability of an event based on the prior knowledge of the conditions that might be related to that event:

$$P(\text{model} \mid \text{data}) = \frac{P(\text{data}|\text{model})\, P(\text{model})}{P(\text{data})}$$

In Bayes' equation, $P(\text{model})$ reflects our prior knowledge about the model. Also, $P(\text{data})$ is the distribution of the data. So we can rewrite as:

$$P(\text{label} \mid \text{text}) = \frac{P(\text{label}|\text{text})\, P(\text{label})}{P(\text{text})}$$

where text --> word_1, word_2 and so on.

BernoulliNB is suitable for discrete data (binary/boolean features) that is what we got previously.

[ ]:
```
from sklearn.naive_bayes import BernoulliNB

model = BernoulliNB()

model.fit(X_train, y_train)

p_train = model.predict(X_train)
p_test = model.predict(X_test)
```

After these two scores are calculated we can compare each other, to check the performance of the classifier.

[ ]:
```
from sklearn.metrics import accuracy_score

acc_train = accuracy_score(y_train, p_train)
acc_test = accuracy_score(y_test, p_test)
```

```
print(f'Train {acc_train}, Test {acc_test}')
```

```
Train 0.9918666666666667, Test 0.87648
```

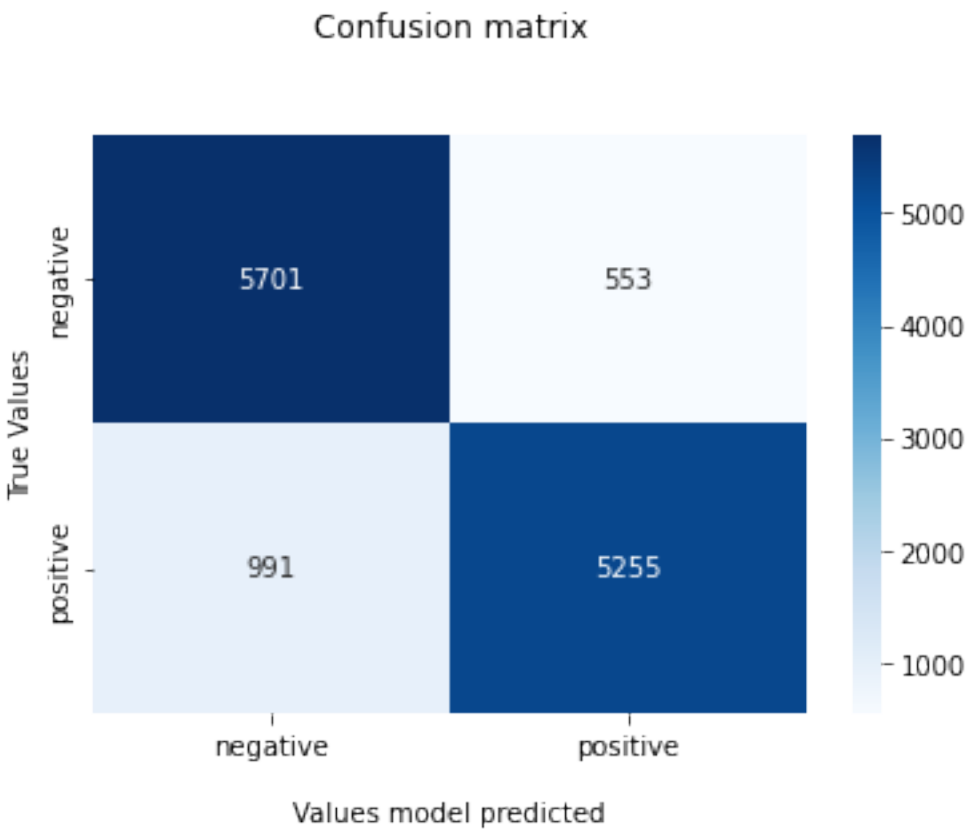
1.0.4 Model Performance Analysis

We'll use two measures: 1. Confusion matrix 2. Classification Report with Precision, Recall and F1-Score.

The confusion matrix is a table that is used to show the number of correct and incorrect predictions on a classification problem when the real values of the Test Set are known. It is of the format

TP	FP
FN	TN

[ ]:
```
confusionMatrix = pd.crosstab(y_test, p_test)
fx = sns.heatmap(confusionMatrix, annot=True, cmap='Blues', fmt='d')
fx.set_title('Confusion matrix\n\n');
fx.set_xlabel('\nValues model predicted')
fx.set_ylabel('True Values')
plt.show()
```



[ ]:
```
from sklearn.metrics import classification_report

classificationReportNotScaled = classification_report(y_test, p_test)
print(f"Classification Report\n{classificationReportNotScaled}")
```

```
Classification Report
              precision    recall  f1-score   support

    negative       0.85      0.91      0.88      6254
    positive       0.90      0.84      0.87      6246

    accuracy                           0.88     12500
   macro avg       0.88      0.88      0.88     12500
weighted avg       0.88      0.88      0.88     12500
```