IMD Sentiment Analysis with NLP

```
In [2]:
          import numpy as np
          import pandas as pd
          import matplotlib.pyplot as plt
          from sklearn.feature extraction.text import CountVectorizer
          from sklearn.ensemble import RandomForestClassifier
          from sklearn.metrics import roc auc score
          from bs4 import BeautifulSoup
          import re
          import nltk
          from sklearn.model_selection import train_test_split
          from nltk.corpus import stopwords
          #load our data set
 In [7]:
          df = pd.read_csv('NLPlabeledData.tsv', delimiter="\t", quoting=3)
In [69]: df.head()
Out[69]:
                   id sentiment
                                                                  review
          0 "5814 8"
                                  "With all this stuff going down at the moment ...
          1 "2381_9"
                                  "\"The Classic War of the Worlds\" by Timothy ...
          2 "7759_3"
                              0
                                  "The film starts with a manager (Nicholas Bell...
          3 "3630 4"
                                 "It must be assumed that those who praised thi...
          4 "9495_8"
                              1 "Superbly trashy and wondrously unpretentious ...
 In [9]:
          len(df)
          25000
 Out[9]:
In [10]:
          len(df["review"])
          25000
Out[10]:
In [11]:
          #To clear stopwords, we need to download the stopwords word set from the
          # We do this with nltk
          nltk.download('stopwords')
          [nltk_data] Downloading package stopwords to
                          /Users/anildemirel/nltk data...
          [nltk data]
          [nltk_data]
                         Package stopwords is already up-to-date!
          True
Out[11]:
```

Data Cleaning Operations

First, we will delete HTML tags from review sentences using the BeautifulSoup module.

To explain how these processes are done, let's first select a single review and see how it is done for you:

```
In [13]: sample_review= df.review[0]
    sample_review
```

Out[13]:

"With all this stuff going down at the moment with MJ i\'ve started list ening to his music, watching the odd documentary here and there, watched The Wiz and watched Moonwalker again. Maybe i just want to get a certain insight into this guy who i thought was really cool in the eighties just to maybe make up my mind whether he is quilty or innocent. Moonwalker is part biography, part feature film which i remember going to see at the ci nema when it was originally released. Some of it has subtle messages abou t MJ\'s feeling towards the press and also the obvious message of drugs a re bad m\'kay.

Visually impressive but of course this is all a bout Michael Jackson so unless you remotely like MJ in anyway then you ar e going to hate this and find it boring. Some may call MJ an egotist for consenting to the making of this movie BUT MJ and most of his fans would say that he made it for the fans which if true is really nice of him.

The actual feature film bit when it finally starts is only on for 20 minutes or so excluding the Smooth Criminal sequence and Joe Pesci is convincing as a psychopathic all powerful drug lord. Why he wants MJ dead so bad is beyond me. Because MJ overheard his plans? Nah, Joe Pesci\'s ch aracter ranted that he wanted people to know it is he who is supplying dr ugs etc so i dunno, maybe he just hates MJ\'s music.

Lots of c ool things in this like MJ turning into a car and a robot and the whole S peed Demon sequence. Also, the director must have had the patience of a s aint when it came to filming the kiddy Bad sequence as usually directors hate working with one kid let alone a whole bunch of them performing a co mplex dance scene.

Bottom line, this movie is for people who l ike MJ on one level or another (which i think is most people). If not, th en stay away. It does try and give off a wholesome message and ironically MJ\'s bestest buddy in this movie is a girl! Michael Jackson is truly one of the most talented people ever to grace this planet but is he guilty? W ell, with all the attention i\'ve gave this subject....hmmm well i don\'t know because people can be different behind closed doors, i know this for a fact. He is either an extremely nice but stupid guy or one of the most sickest liars. I hope he is not the latter."'

```
In [14]: # After cleaning the HTML tags..
sample_review = BeautifulSoup(sample_review).get_text()
sample_review
```

Out[14]:

'"With all this stuff going down at the moment with MJ i\'ve started list ening to his music, watching the odd documentary here and there, watched The Wiz and watched Moonwalker again. Maybe i just want to get a certain insight into this guy who i thought was really cool in the eighties just to maybe make up my mind whether he is quilty or innocent. Moonwalker is part biography, part feature film which i remember going to see at the ci nema when it was originally released. Some of it has subtle messages abou t MJ\'s feeling towards the press and also the obvious message of drugs a re bad m\'kay.Visually impressive but of course this is all about Michael Jackson so unless you remotely like MJ in anyway then you are going to ha te this and find it boring. Some may call MJ an egotist for consenting to the making of this movie BUT MJ and most of his fans would say that he ma de it for the fans which if true is really nice of him. The actual feature film bit when it finally starts is only on for 20 minutes or so excluding the Smooth Criminal sequence and Joe Pesci is convincing as a psychopathi c all powerful drug lord. Why he wants MJ dead so bad is beyond me. Becau se MJ overheard his plans? Nah, Joe Pesci\'s character ranted that he wan ted people to know it is he who is supplying drugs etc so i dunno, maybe he just hates MJ\'s music.Lots of cool things in this like MJ turning int o a car and a robot and the whole Speed Demon sequence. Also, the directo r must have had the patience of a saint when it came to filming the kiddy Bad sequence as usually directors hate working with one kid let alone a w hole bunch of them performing a complex dance scene. Bottom line, this mov ie is for people who like MJ on one level or another (which i think is mo st people). If not, then stay away. It does try and give off a wholesome message and ironically MJ\'s bestest buddy in this movie is a girl! Micha el Jackson is truly one of the most talented people ever to grace this pl anet but is he quilty? Well, with all the attention i\'ve gave this subje ct....hmmm well i don\'t know because people can be different behind clos ed doors, i know this for a fact. He is either an extremely nice but stup id guy or one of the most sickest liars. I hope he is not the latter."'

```
In [15]: # we clean it from punctuation and numbers - using regex..
sample_review = re.sub("[^a-zA-Z]",' ',sample_review)
sample_review
```

Out[15]:

' With all this stuff going down at the moment with MJ i ve started liste ning to his music watching the odd documentary here and there watched T he Wiz and watched Moonwalker again Maybe i just want to get a certain i nsight into this guy who i thought was really cool in the eighties just t o maybe make up my mind whether he is quilty or innocent Moonwalker is p art biography part feature film which i remember going to see at the cin ema when it was originally released Some of it has subtle messages about MJ s feeling towards the press and also the obvious message of drugs are bad m kay Visually impressive but of course this is all about Michael Jac kson so unless you remotely like MJ in anyway then you are going to hate this and find it boring Some may call MJ an egotist for consenting to th e making of this movie BUT MJ and most of his fans would say that he made it for the fans which if true is really nice of him The actual feature fi lm bit when it finally starts is only on for minutes or so excluding t he Smooth Criminal sequence and Joe Pesci is convincing as a psychopathic all powerful drug lord Why he wants MJ dead so bad is beyond me Because MJ overheard his plans Nah Joe Pesci s character ranted that he wanted people to know it is he who is supplying drugs etc so i dunno maybe he j ust hates MJ s music Lots of cool things in this like MJ turning into a c ar and a robot and the whole Speed Demon sequence Also the director mus t have had the patience of a saint when it came to filming the kiddy Bad sequence as usually directors hate working with one kid let alone a whole bunch of them performing a complex dance scene Bottom line this movie is for people who like MJ on one level or another which i think is most peo If not then stay away It does try and give off a wholesome messag e and ironically MJ s bestest buddy in this movie is a girl Michael Jack son is truly one of the most talented people ever to grace this planet bu t is he guilty Well with all the attention i ve gave this subject mm well i don t know because people can be different behind closed doors i know this for a fact He is either an extremely nice but stupid guy or one of the most sickest liars I hope he is not the latter

In [16]: # convert to lowercase, We do this so that machine learning algorithms do
start with a capital letter as different words
sample_review = sample_review.lower()
sample_review

Out[16]:

' with all this stuff going down at the moment with mj i ve started liste ning to his music watching the odd documentary here and there watched t he wiz and watched moonwalker again maybe i just want to get a certain i nsight into this guy who i thought was really cool in the eighties just t o maybe make up my mind whether he is quilty or innocent moonwalker is p art biography part feature film which i remember going to see at the cin ema when it was originally released some of it has subtle messages about mj s feeling towards the press and also the obvious message of drugs are bad m kay visually impressive but of course this is all about michael jac kson so unless you remotely like mj in anyway then you are going to hate this and find it boring some may call mj an egotist for consenting to th e making of this movie but mj and most of his fans would say that he made it for the fans which if true is really nice of him the actual feature fi lm bit when it finally starts is only on for minutes or so excluding t he smooth criminal sequence and joe pesci is convincing as a psychopathic all powerful drug lord why he wants mj dead so bad is beyond me because mj overheard his plans nah joe pesci s character ranted that he wanted people to know it is he who is supplying drugs etc so i dunno maybe he j ust hates mj s music lots of cool things in this like mj turning into a c ar and a robot and the whole speed demon sequence also the director mus t have had the patience of a saint when it came to filming the kiddy bad sequence as usually directors hate working with one kid let alone a whole bunch of them performing a complex dance scene bottom line this movie is for people who like mj on one level or another which i think is most peo if not then stay away it does try and give off a wholesome messag e and ironically mj s bestest buddy in this movie is a girl michael jack son is truly one of the most talented people ever to grace this planet bu t is he guilty well with all the attention i ve gave this subject mm well i don t know because people can be different behind closed doors i know this for a fact he is either an extremely nice but stupid guy or one of the most sickest liars i hope he is not the latter

```
In [17]: # stopwords
          # First, we split the words with split and convert them to a list. our go
          sample review = sample review.split()
In [18]:
          sample review
Out[18]: ['with',
           'all',
           'this',
           'stuff',
           'going',
           'down',
           'at',
           'the',
           'moment',
           'with',
           'mj',
           'i',
           've',
           'started',
           'listening',
           'to',
           'his',
           'music',
           'watching',
```

```
'the',
'odd',
'documentary',
'here',
'and',
'there',
'watched',
'the',
'wiz',
'and',
'watched',
'moonwalker',
'again',
'maybe',
'i',
'just',
'want',
'to',
'get',
'a',
'certain',
'insight',
'into',
'this',
'guy',
'who',
'i',
'thought',
'was',
'really',
'cool',
'in',
'the',
'eighties',
'just',
'to',
'maybe',
'make',
'up',
'my',
'mind',
'whether',
'he',
'is',
'guilty',
'or',
'innocent',
'moonwalker',
'is',
'part',
'biography',
'part',
'feature',
'film',
'which',
'i',
'remember',
```

```
'going',
'to',
'see',
'at',
'the',
'cinema',
'when',
'it',
'was',
'originally',
'released',
'some',
'of',
'it',
'has',
'subtle',
'messages',
'about',
'mj',
's',
'feeling',
'towards',
'the',
'press',
'and',
'also',
'the',
'obvious',
'message',
'of',
'drugs',
'are',
'bad',
'm',
'kay',
'visually',
'impressive',
'but',
'of',
'course',
'this',
'is',
'all',
'about',
'michael',
'jackson',
'so',
'unless',
'you',
'remotely',
'like',
'mj',
'in',
'anyway',
'then',
'you',
'are',
```

```
'going',
'to',
'hate',
'this',
'and',
'find',
'it',
'boring',
'some',
'may',
'call',
'mj',
'an',
'egotist',
'for',
'consenting',
'to',
'the',
'making',
'of',
'this',
'movie',
'but',
'mj',
'and',
'most',
'of',
'his',
'fans',
'would',
'say',
'that',
'he',
'made',
'it',
'for',
'the',
'fans',
'which',
'if',
'true',
'is',
'really',
'nice',
'of',
'him',
'the',
'actual',
'feature',
'film',
'bit',
'when',
'it',
'finally',
'starts',
'is',
'only',
```

```
'on',
'for',
'minutes',
'or',
'so',
'excluding',
'the',
'smooth',
'criminal',
'sequence',
'and',
'joe',
'pesci',
'is',
'convincing',
'as',
'a',
'psychopathic',
'all',
'powerful',
'drug',
'lord',
'why',
'he',
'wants',
'mj',
'dead',
'so',
'bad',
'is',
'beyond',
'me',
'because',
'mj',
'overheard',
'his',
'plans',
'nah',
'joe',
'pesci',
's',
'character',
'ranted',
'that',
'he',
'wanted',
'people',
'to',
'know',
'it',
'is',
'he',
'who',
'is',
'supplying',
'drugs',
'etc',
```

```
'so',
'i',
'dunno',
'maybe',
'he',
'just',
'hates',
'mj',
's',
'music',
'lots',
'of',
'cool',
'things',
'in',
'this',
'like',
'mj',
'turning',
'into',
'a',
'car',
'and',
'a',
'robot',
'and',
'the',
'whole',
'speed',
'demon',
'sequence',
'also',
'the',
'director',
'must',
'have',
'had',
'the',
'patience',
'of',
'a',
'saint',
'when',
'it',
'came',
'to',
'filming',
'the',
'kiddy',
'bad',
'sequence',
'as',
'usually',
'directors',
'hate',
'working',
'with',
```

```
'one',
'kid',
'let',
'alone',
'a',
'whole',
'bunch',
'of',
'them',
'performing',
'a',
'complex',
'dance',
'scene',
'bottom',
'line',
'this',
'movie',
'is',
'for',
'people',
'who',
'like',
'mj',
'on',
'one',
'level',
'or',
'another',
'which',
'i',
'think',
'is',
'most',
'people',
'if',
'not',
'then',
'stay',
'away',
'it',
'does',
'try',
'and',
'give',
'off',
'a',
'wholesome',
'message',
'and',
'ironically',
'mj',
's',
'bestest',
'buddy',
'in',
'this',
```

```
'movie',
'is',
'a',
'girl',
'michael',
'jackson',
'is',
'truly',
'one',
'of',
'the',
'most',
'talented',
'people',
'ever',
'to',
'grace',
'this',
'planet',
'but',
'is',
'he',
'guilty',
'well',
'with',
'all',
'the',
'attention',
'i',
've',
'gave',
'this',
'subject',
'hmmm',
'well',
'i',
'don',
't',
'know',
'because',
'people',
'can',
'be',
'different',
'behind',
'closed',
'doors',
'i',
'know',
'this',
'for',
'a',
'fact',
'he',
'is',
'either',
'an',
```

```
'extremely',
           'nice',
           'but',
           'stupid',
           'guy',
           'or',
           'one',
           'of',
           'the',
           'most',
           'sickest',
           'liars',
           'i',
           'hope',
           'he',
           'is',
           'not',
           'the',
           'latter']
In [19]:
          len(sample review)
          437
Out[19]:
In [20]:
          # sample review without stopwords
          swords = set(stopwords.words("english"))
                                                                              # conversio
          sample review = [w for w in sample review if w not in swords]
          sample review
          ['stuff',
Out[20]:
           'going',
           'moment',
           'mj',
           'started',
           'listening',
           'music',
           'watching',
           'odd',
           'documentary',
           'watched',
           'wiz',
           'watched',
           'moonwalker',
           'maybe',
           'want',
           'get',
           'certain',
           'insight',
           'guy',
           'thought',
           'really',
           'cool',
           'eighties',
           'maybe',
           'make',
           'mind',
           'whether',
```

```
'guilty',
'innocent',
'moonwalker',
'part',
'biography',
'part',
'feature',
'film',
'remember',
'going',
'see',
'cinema',
'originally',
'released',
'subtle',
'messages',
'mj',
'feeling',
'towards',
'press',
'also',
'obvious',
'message',
'drugs',
'bad',
'kay',
'visually',
'impressive',
'course',
'michael',
'jackson',
'unless',
'remotely',
'like',
'mj',
'anyway',
'going',
'hate',
'find',
'boring',
'may',
'call',
'mj',
'egotist',
'consenting',
'making',
'movie',
'mj',
'fans',
'would',
'say',
'made',
'fans',
'true',
'really',
'nice',
'actual',
```

```
'feature',
'film',
'bit',
'finally',
'starts',
'minutes',
'excluding',
'smooth',
'criminal',
'sequence',
'joe',
'pesci',
'convincing',
'psychopathic',
'powerful',
'drug',
'lord',
'wants',
'mj',
'dead',
'bad',
'beyond',
'mj',
'overheard',
'plans',
'nah',
'joe',
'pesci',
'character',
'ranted',
'wanted',
'people',
'know',
'supplying',
'drugs',
'etc',
'dunno',
'maybe',
'hates',
'mj',
'music',
'lots',
'cool',
'things',
'like',
'mj',
'turning',
'car',
'robot',
'whole',
'speed',
'demon',
'sequence',
'also',
'director',
'must',
'patience',
```

```
'saint',
'came',
'filming',
'kiddy',
'bad',
'sequence',
'usually',
'directors',
'hate',
'working',
'one',
'kid',
'let',
'alone',
'whole',
'bunch',
'performing',
'complex',
'dance',
'scene',
'bottom',
'line',
'movie'
'people',
'like',
'mj',
'one',
'level',
'another',
'think',
'people',
'stay',
'away',
'try',
'give',
'wholesome',
'message',
'ironically',
'mj',
'bestest',
'buddy',
'movie',
'girl',
'michael',
'jackson',
'truly',
'one',
'talented',
'people',
'ever',
'grace',
'planet',
'guilty',
'well',
'attention',
'gave',
'subject',
```

```
'hmmm',
           'well',
           'know',
           'people',
           'different',
           'behind',
           'closed',
           'doors',
           'know',
           'fact',
           'either',
           'extremely',
           'nice',
           'stupid',
           'guy',
           'one',
           'sickest',
           'liars',
           'hope',
           'latter']
In [21]:
         len(sample review)
         219
Out[21]:
In [22]:
         # After describing the cleanup process, we now batch clean the reviews in
         # for this purpose we first create a function:
         def process(review):
In [23]:
              # review without HTML tags
             review = BeautifulSoup(review).get_text()
              # review without punctuation and numbers
             review = re.sub("[^a-zA-Z]",' ',review)
              # converting into lowercase and splitting to eliminate stopwords
             review = review.lower()
              review = review.split()
              # review without stopwords
              swords = set(stopwords.words("english"))
                                                                              # conve
              review = [w for w in review if w not in swords]
              # splitted paragraph'ları space ile birleştiriyoruz return
              return(" ".join(review))
In [24]: # We clean our training data with the help of the above function:
         # We can see the status of the review process by printing a line after ev
         train x tum = []
          for r in range(len(df["review"])):
              if (r+1)%1000 == 0:
                  print("No of reviews processed =", r+1)
              train_x_tum.append(process(df["review"][r]))
         /Users/anildemirel/opt/anaconda3/lib/python3.9/site-packages/bs4/ init
          .py:435: MarkupResemblesLocatorWarning: The input looks more like a filen
         ame than markup. You may want to open this file and pass the filehandle i
         nto Beautiful Soup.
           warnings.warn(
```

```
No of reviews processed = 1000
No of reviews processed = 2000
No of reviews processed = 3000
No of reviews processed = 4000
No of reviews processed = 5000
No of reviews processed = 6000
No of reviews processed = 7000
No of reviews processed = 8000
No of reviews processed = 9000
No of reviews processed = 10000
No of reviews processed = 11000
No of reviews processed = 12000
No of reviews processed = 13000
No of reviews processed = 14000
No of reviews processed = 15000
No of reviews processed = 16000
No of reviews processed = 17000
No of reviews processed = 18000
No of reviews processed = 19000
No of reviews processed = 20000
No of reviews processed = 21000
No of reviews processed = 22000
No of reviews processed = 23000
No of reviews processed = 24000
No of reviews processed = 25000
```

Train, test split

```
In [27]: x = train_x_tum
y = np.array(df["sentiment"])

# train test split
train_x, test_x, y_train, y_test = train_test_split(x,y, test_size = 0.1)
```

Create Bag of Words

```
In [28]:
         # Using the count vectorizer function in sklearn, we create a bag of word
         vectorizer = CountVectorizer( max_features = 5000 )
         # we convert our train data to feature vector matrix
         train x = vectorizer.fit transform(train x)
In [29]:
         train x
         <22500x5000 sparse matrix of type '<class 'numpy.int64'>'
Out[29]:
                 with 1779382 stored elements in Compressed Sparse Row format>
In [30]:
         # We're converting it to an array because it requires an array for the fi
         train_x = train_x.toarray()
         train_y = y_train
In [31]:
         train_x.shape, train_y.shape
```

```
Out[31]: ((22500, 5000), (22500,))
In [32]: train_y
Out[32]: array([1, 0, 0, ..., 0, 0])
```

We create Random Forest Model and fit

Now it's time for our test data...

Prediction

```
In [45]:
          print(test_xx[4])
         [0 0 0 ... 1 0 0]
In [46]:
          print(test predict[4])
In [47]: print(test_xx[2])
         [0 0 0 ... 0 0 0]
In [48]:
         print(test_predict[2])
In [54]: print(test_xx[4])
         [0 0 0 ... 1 0 0]
In [55]:
         print(test_predict[4])
         0
In [56]: print(test_xx[3])
         [0 0 0 ... 0 0 0]
         print(test_predict[3])
In [57]:
In [87]: print(test_xx[20])
         [0 0 0 ... 0 0 0]
In [88]:
         print(test_predict[20])
         1
In [81]: print(test xx[17])
         [0 0 0 ... 0 0 0]
In [82]:
         print(test predict[17])
         0
In [77]: print(test_xx[12])
         [0 0 0 ... 0 0 0]
In [78]:
         print(test_predict[12])
In [64]: print(test_xx[7])
         [0 0 0 ... 0 0 0]
```

In [65]: print(test_predict[7])

0

In [70]: df.head(24)

Out[70]:

III [70]1 azviloda(21)

	id	sentiment	review
0	"5814_8"	1	"With all this stuff going down at the moment
1	"2381_9"	1	"\"The Classic War of the Worlds\" by Timothy
2	"7759_3"	0	"The film starts with a manager (Nicholas Bell
3	"3630_4"	0	"It must be assumed that those who praised thi
4	"9495_8"	1	"Superbly trashy and wondrously unpretentious
5	"8196_8"	1	"I dont know why people think this is such a b
6	"7166_2"	0	"This movie could have been very good, but com
7	"10633_1"	0	"I watched this video at a friend's house. I'm
8	"319_1"	0	"A friend of mine bought this film for £1, and
9	"8713_10"	1	" This movie is full of references
10	"2486_3"	0	"What happens when an army of wetbacks, towelh
11	"6811_10"	1	"Although I generally do not like remakes beli
12	"11744_9"	1	"\"Mr. Harvey Lights a Candle\" is anchored by
13	"7369_1"	0	"I had a feeling that after \"Submerged\", thi
14	"12081_1"	0	"note to George Litman, and others: the Myster
15	"3561_4"	0	"Stephen King adaptation (scripted by King him
16	"4489_1"	0	"`The Matrix' was an exciting summer blockbust
17	"3951_2"	0	"Ulli Lommel's 1980 film 'The Boogey Man' is n
18	"3304_10"	1	"This movie is one among the very few Indian m
19	"9352_10"	1	"Most people, especially young people, may not
20	"3374_7"	1	"\"Soylent Green\" is one of the best and most
21	"10782_7"	1	"Michael Stearns plays Mike, a sexually frustr
22	"5414_10"	1	"This happy-go-luck 1939 military swashbuckler
23	"10492_1"	0	"I would love to have that two hours of my lif

In []: