UCI

Drug Review Dataset

In [ ]:

```
import pandas as pd
```

In [2]:

```
data_train = pd.read_csv('.....\\drugsCom_raw\\drugsComTrain_raw.tsv',delimiter='\t')
data_test = pd.read_csv('......\\drugsCom_raw\\drugsComTest_raw.tsv' ,delimiter='\t')
```

In [ ]:

```
```

In [3]:

```
df = pd.concat([data_train,data_test])  # combine the two dataFrames into one for a bigger data size and ease of preprocessing
```

In [4]:

```
data_train.shape
```

Out[4]:

```
(161297, 7)
```

In [5]:

```
data_test.shape
```

Out[5]:

```
(53766, 7)
```

In [6]:

```
df.head()
```

Out[6]:

	Unnamed: 0	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9.0	May 20, 2012	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8.0	April 27, 2010	192
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5.0	December 14, 2009	17
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8.0	November 3, 2015	10
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9.0	November 27, 2016	37

In [7]:

```
df.columns = ['Id','drugName','condition','review','rating','date','usefulCount']    #rename columns
```

In [8]:

```
df.head()
```

Out[8]:

	Id	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9.0	May 20, 2012	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8.0	April 27, 2010	192
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5.0	December 14, 2009	17
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8.0	November 3, 2015	10
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9.0	November 27, 2016	37

In [9]:

```
df['date'] = pd.to_datetime(df['date'])    #convert date to datetime eventhough we are not using date in this
```

In [10]:

```
df['date'].head()             #confirm conversion
```

Out[10]:

```
0   2012-05-20
1   2010-04-27
2   2009-12-14
3   2015-11-03
4   2016-11-27
Name: date, dtype: datetime64[ns]
```

In [11]:

```
df2 = df[['Id','review','rating']].copy()    # create a new dataframe with just review and rating for sentiment analysis
```

In [12]:

```
df.head()             #confirm conversion
```

Out[12]:

	Id	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9.0	2012-05-20	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8.0	2010-04-27	192
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5.0	2009-12-14	17
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8.0	2015-11-03	10
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9.0	2016-11-27	37

In [13]:

```
df2.head()
```

Out[13]:

	Id	review	rating
0	206461	"It has no side effect, I take it in combinati...	9.0
1	95260	"My son is halfway through his fourth week of ...	8.0
2	92703	"I used to take another oral contraceptive, wh...	5.0
3	138000	"This is my first time using any form of birth...	8.0
4	35696	"Suboxone has completely turned my life around...	9.0

In [14]:

```
df2.isnull().any().any()    # check for null
```

Out[14]:

```
False
```

In [15]:

```
df2.info(null_counts=True)          #another way to check for null
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 215063 entries, 0 to 53765
Data columns (total 3 columns):
 #   Column   Non-Null Count   Dtype  
---  ------   --------------   -----  
 0   Id       215063 non-null  int64  
 1   review   215063 non-null  object 
 2   rating   215063 non-null  float64
dtypes: float64(1), int64(1), object(1)
memory usage: 6.6+ MB
```

In [16]:

```
df2.info()       #check for datatype, also shows null
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 215063 entries, 0 to 53765
Data columns (total 3 columns):
 #   Column   Non-Null Count   Dtype  
---  ------   --------------   -----  
 0   Id       215063 non-null  int64  
 1   review   215063 non-null  object 
 2   rating   215063 non-null  float64
dtypes: float64(1), int64(1), object(1)
memory usage: 6.6+ MB
```

In [17]:

```
df2['Id'].unique()       # shows unique Id as array
```

Out[17]:

```
array([206461,  95260,  92703, ..., 130945,  47656, 113712], dtype=int64)
```

In [18]:

```
df2['Id'].count()      #count total number of items in the Id column
```

Out[18]:

```
215063
```

In [19]:

```
df2['Id'].nunique()      #shows unique Id values
```

Out[19]:

```
215063
```

```
In [20]: df['review'][1]          # access indivdual value
```

```
Out[20]: 1    "My son is halfway through his fourth week of ...
         1    "My son has Crohn&#039;s disease and has done ...
         Name: review, dtype: object
```

```
In [21]: df.review[1]             # another method to assess individual value in a Serie
         s
```

```
Out[21]: 1    "My son is halfway through his fourth week of ...
         1    "My son has Crohn&#039;s disease and has done ...
         Name: review, dtype: object
```

```
In [22]: import nltk
         nltk.download(['punkt','stopwords'])
```

```
[nltk_data] Downloading package punkt to C:\Users\PC-
[nltk_data]     Tiger\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to C:\Users\PC-
[nltk_data]     Tiger\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

```
Out[22]: True
```

```
In [23]: from nltk.corpus import stopwords
         stopwords = stopwords.words('english')
```

```
In [24]: df2['cleanReview'] = df2['review'].apply(lambda x: ' '.join([item for item in
         x.split() if item not in stopwords]))     # remove stopwords from review
```

```
In [26]: df2['cleanReview'] = df2['review'].apply(lambda x: ' '.join([item for item in
         x.split() if item not in stopwords]))     # remove stopwords from review
```

```
In [56]: import vaderSentiment
         from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
         analyzer = SentimentIntensityAnalyzer()
```

```
In [57]: df2['vaderReviewScore'] = df2['cleanReview'].apply(lambda x: analyzer.polarity
         _scores(x)['compound'])
```

```
In [59]: positive_num = len(df2[df2['vaderReviewScore'] >=0.05])
         neutral_num = len(df2[(df2['vaderReviewScore'] >-0.05) & (df2['vaderReviewScor
         e']<0.05)])
         negative_num = len(df2[df2['vaderReviewScore']<=-0.05])
```

```
In [60]: positive_num,neutral_num, negative_num
```

```
Out[60]: (106198, 9035, 99830)
```

```
In [61]: df2['vaderSentiment']= df2['vaderReviewScore'].map(lambda x:int(2) if x>=0.05
         else int(1) if x<=-0.05 else int(0) )
```

In [62]:
```
df2['vaderSentiment'].value_counts()
```

Out[62]:
```
2    106198
1     99830
0      9035
Name: vaderSentiment, dtype: int64
```

In [63]:
```
Total_vaderSentiment = positive_num + neutral_num + negative_num
Total_vaderSentiment
```

Out[63]:
```
215063
```

In [64]:
```
df2.loc[df2['vaderReviewScore'] >=0.05,"vaderSentimentLabel"] ="positive"
df2.loc[(df2['vaderReviewScore'] >-0.05) & (df2['vaderReviewScore']<0.05),"vaderSentimentLabel"]= "neutral"
df2.loc[df2['vaderReviewScore']<=-0.05,"vaderSentimentLabel"] = "negative"
```

In [65]:
```
df2.shape
```

Out[65]:
```
(215063, 9)
```

In [66]:
```
positive_rating = len(df2[df2['rating'] >=7.0])
neutral_rating = len(df2[(df2['rating'] >=4) & (df2['rating']<7)])
negative_rating = len(df2[df2['rating']<=3])
```

In [67]:
```
positive_rating,neutral_rating,negative_rating
```

Out[67]:
```
(142306, 25856, 46901)
```

In [68]:
```
Total_rating = positive_rating+neutral_rating+negative_rating
Total_rating
```

Out[68]:
```
215063
```

In [69]:
```
df2['ratingSentiment']= df2['rating'].map(lambda x:int(2) if x>=7 else int(1) if x<=3 else int(0) )
```

In [70]:
```
df2['ratingSentiment'].value_counts()
```

Out[70]:
```
2    142306
1     46901
0     25856
Name: ratingSentiment, dtype: int64
```

In [72]:
```
df2.loc[df2['rating'] >=7.0,"ratingSentimentLabel"] ="positive"
df2.loc[(df2['rating'] >=4.0) & (df2['rating']<7.0),"ratingSentimentLabel"]= "neutral"
df2.loc[df2['rating']<=3.0,"ratingSentimentLabel"] = "negative"
```

In [98]:
```
df2 = df2[['Id','review','cleanReview','rating','ratingSentiment','ratingSentimentLabel','vaderReviewScore','vaderSentimentLabel','vaderSentiment']]
```

=============================

In [104]:
```
data_df=df2.drop(['review','cleanReview'],axis=1)
```

In [149]:
```
data_df.head()
```

Out[149]:

	Id	review	cleanReview	rating	ratingSentiment	vaderReviewScore	vaderSentimer
0	206461	"It has no side effect, I take it in combinati...	"It side effect, I take combination Bystolic 5...	9.0	2	0.0000	
1	95260	"My son is halfway through his fourth week of ...	"My son halfway fourth week Intuniv. We became...	8.0	2	0.9070	
2	92703	"I used to take another oral contraceptive, wh...	"I used take another oral contraceptive, 21 pi...	5.0	0	0.7096	
3	138000	"This is my first time using any form of birth...	"This first time using form birth control. I&#...	8.0	2	0.7184	
4	35696	"Suboxone has completely turned my life around...	"Suboxone completely turned life around. I fee...	9.0	2	0.9403	

In [150]:
```
data_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 215063 entries, 0 to 53765
Data columns (total 8 columns):
 #   Column               Non-Null Count   Dtype  
---  ------               --------------   -----  
 0   Id                   215063 non-null  int64  
 1   review               215063 non-null  object 
 2   cleanReview          215063 non-null  object 
 3   rating               215063 non-null  float64
 4   ratingSentiment      215063 non-null  int64  
 5   vaderReviewScore     215063 non-null  float64
 6   vaderSentimentLabel  215063 non-null  object 
 7   vaderSentiment       215063 non-null  int64  
dtypes: float64(2), int64(3), object(3)
memory usage: 14.8+ MB
```

In [145]:
```
#data_df=df2.drop(['ratingSentimentLabel'],axis=1)
```

In [169]:

```python
from sklearn.preprocessing import LabelEncoder
```

In [188]:

```python
encoder = LabelEncoder()
data_cat = data_df["review"]
data_cat_encod = encoder.fit_transform(data_cat)
data_cat_encod = pd.DataFrame(data_cat_encod,columns=["review"])
data_cat_encod.head()
```

Out[188]:

	review
0	88969
1	98512
2	66084
3	113366
4	107807

In [152]:

```python
encoder = LabelEncoder()
data_cat = data_df["cleanReview"]
data_cat_encod = encoder.fit_transform(data_cat)
data_cat_encod = pd.DataFrame(data_cat_encod,columns=["cleanReview"])
data_cat_encod.head()
```

Out[152]:

	cleanReview
0	89755
1	98510
2	72340
3	113308
4	107788

In [153]:

```python
encoder = LabelEncoder()
data_cat = data_df["vaderSentimentLabel"]
data_cat_encod = encoder.fit_transform(data_cat)
data_cat_encod = pd.DataFrame(data_cat_encod,columns=["vaderSentimentLabel"])
data_cat_encod.head()
```

Out[153]:

	vaderSentimentLabel
0	1
1	2
2	2
3	2
4	2

In [189]:
```
encoder
```

Out[189]:
```
LabelEncoder()
```

In [148]:
```
#df2.to_csv('processed.csv')    # To save preprocessed dataset to csv
```

In [ ]:
```

```

In [ ]:
```

```

In [78]:
```
import os
#os.stat('processed.csv').st_size         # Check size of csv file About 181MB
```

Out[78]:
```
181826800
```

In [ ]:
```

```

In [79]:
```
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 215063 entries, 0 to 53765
Data columns (total 9 columns):
 #   Column                Non-Null Count   Dtype  
---  ------                --------------   -----  
 0   Id                    215063 non-null  int64  
 1   review                215063 non-null  object 
 2   cleanReview           215063 non-null  object 
 3   rating                215063 non-null  float64
 4   ratingSentiment       215063 non-null  int64  
 5   ratingSentimentLabel  215063 non-null  object 
 6   vaderReviewScore      215063 non-null  float64
 7   vaderSentiment        215063 non-null  int64  
 8   vaderSentimentLabel   215063 non-null  object 
dtypes: float64(2), int64(3), object(4)
memory usage: 16.4+ MB
```

In [80]:
```
#df2.to_csv('processed.csv.gz',compression='gzip')
```

In [53]:
```
#os.stat('processed.csv.gz').st_size    #compressed to about 54MB
```

Out[53]:
```
54014522
```

In [54]:

```
df2.head()
```

Out[54]:

	Id	review	cleanReview	rating	ratingSentiment	ratingSentimentLabel	vaderRevie
0	206461	"It has no side effect, I take it in combinati...	"It side effect, I take combination Bystolic 5...	9.0	2	positive	
1	95260	"My son is halfway through his fourth week of ...	"My son halfway fourth week Intuniv. We became...	8.0	2	positive	
2	92703	"I used to take another oral contraceptive, wh...	"I used take another oral contraceptive, 21 pi...	5.0	0	neutral	
3	138000	"This is my first time using any form of birth...	"This first time using form birth control. I&#...	8.0	2	positive	
4	35696	"Suboxone has completely turned my life around...	"Suboxone completely turned life around. I fee...	9.0	2	positive	

In [ ]:

```
dfcopy = df2.copy()
```