Ratings Prediction

Submitted by:

Abhishek Jain

ACKNOWLEDGMENT

I take great pleasure to thank and acknowledge the help provided by **Flip Robo Technologies**. I extend whole hearted thanks to Mrs. Khushboo Garg who become my Mentor and with whom I worked and learned a lot and for enlightening me with her knowledge and experience to grow with the corporate working. Her guidance at every stage of the Project enabled me to successfully complete this Project which otherwise would not have been possible without her consent encouragement and motivation. Without the support it was not possible for me to complete the report with fullest endeavour.

INTRODUCTION

- Objective of problem

  Review Rating Prediction attempts to infer from the review's content the user's numerical rating (often between 1 and 5 stars). Helping website visitors determine the rating of their reviews is a good usage of the Rating Prediction job. Recognize phoney or dubious online reviews.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Data Preliminary data analysis must be performed to gain a deeper understanding of the quality of the data, in terms of outliers and the skewedness of the figures, descriptive statistics, and other factors. Understanding and preparation are essential parts of building a model because they provide insight into the data and what corrections or modifications shall be made before designing and executing the model. To do that, category and numerical variables were statistically analysed. Additionally, it helps to be aware of the key factors that influence how prices are determined. This was accomplished by creating a correlation matrix for each attribute to comprehend the relationships between the various components.

- Data Sources and their formats

  The project deals with Indian ecommerce website. Using Selenium, the dataset from flipkart.com and amazon was scraped in order to build the effective intelligent model.

In [5]:
```
# Import dataset

df=pd.read_csv('review_comments_data.csv', index_col=0)
df.head(2)
```

Out[5]:

	Title	Review	Rating
0	Terrible product	Its only 7 months I bought this product. It wa...	1
1	Absolute rubbish!	Automatically Disconnected so many times.	1

In [6]:
```
print("Dataset have ",df.shape[0] , 'rows and ', df.shape[1] ,'columns')
```

```
Dataset have  32431 rows and  3 columns
```

In [7]:
```
# We have 32,081 records and 3 features
```

In [8]:
```
df['Rating'].value_counts(normalize=True)
```

Out[8]:
```
5    0.352348
1    0.253739
4    0.220561
3    0.104190
2    0.069162
Name: Rating, dtype: float64
```

3 Features have been scrapped.

A. Title
B. Review
C. Rating

- Hardware and Software Requirements and Tools Used

  Hardware:

  Software: Latest Anaconda for Jupyter

  Python Libraries:

  Pandas , Numpy, seaborn, matplotlib, scikit-learn,

Model/s Development and Evaluation

- Identification of possible problem-solving approaches
  Used NLP for text pre-processing as our data contains emojis,numerics,spaces etc.

- Testing of Identified Approaches (Algorithms)

1. Remove all email addresses
2. Remove all website links if any
3. Capture emojis
4. Remove all special character
5. Convert into lower case
6. Stemmer/Lemmitizer to convert into base word

Then use machine learning tools to find out the best model.

- Run and Evaluate selected models
  Logistic Regression
  Multinomial NB
  Decision Tree Classifier
  SVC

- Key Metrics for success in solving problem under consideration
  All the data is first converted into string and then data processing is done on it.

Since the dataset we obtained is irregular , having unequal number of similar data according to the ratings(1,2,3,4,5) so we have to use SMOTE technique to balance data.

Visualizations

[10]:
```
# FIrst of all, we will remove duplicate entries which will not be usful for any prediction other than bias the model

df.drop_duplicates(inplace=True)
```

[11]:
```
print("Dataset after removing duplicate entries is  ",df.shape[0] , 'rows and ', df.shape[1] ,'columns')
```

```
Dataset after removing duplicate entries is   24267 rows and  3 columns
```

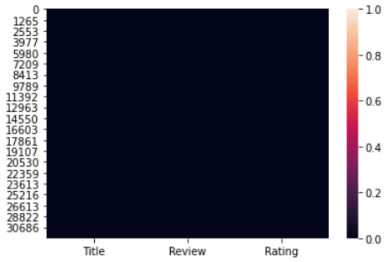
[12]:
```
df
```

t[12]:

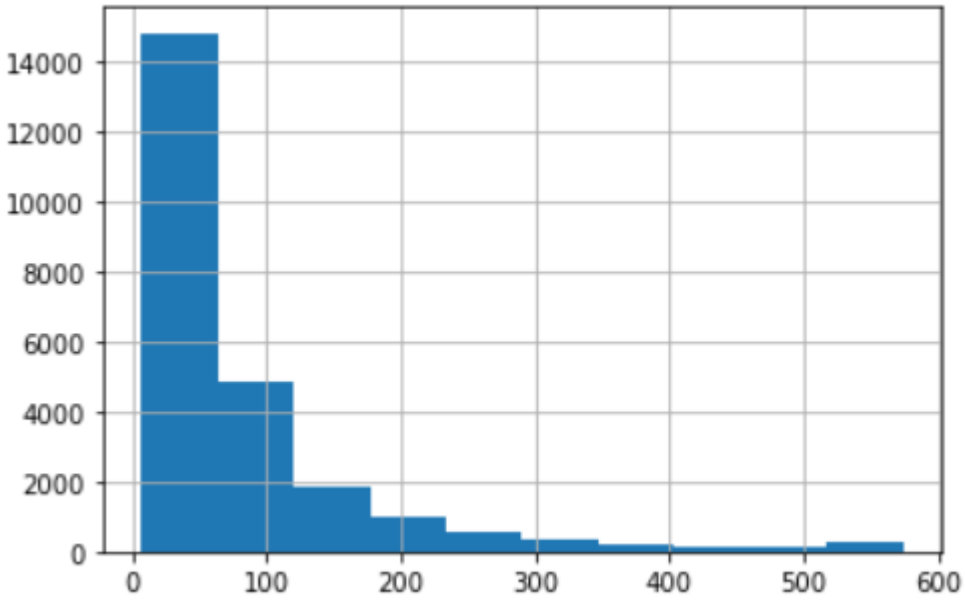
	Title	Review	Rating
0	Terrible product	Its only 7 months I bought this product. It wa...	1
1	Absolute rubbish!	Automatically Disconnected so many times.	1
2	Worst experience ever!	everything is good expect built quality.....b...	1
3	Worst experience ever!	Worst Product Quality Ever,,It just stop worki...	1
4	Very poor	Bakwas flipkart	1
...	...	...	...
32424	Just okay	I love it 😍😍	3
32425	Decent product	Volume not expected	3
32428	Good	I paid extra for stand. Because no hanging arr...	3
32429	Nice	Damage project	3
32430	Just okay	Product is good in this price...	3

24267 rows × 3 columns

Out[15]: <AxesSubplot:>



We have only 1 null value in Title feature which is difficult to see in graph. As the missing value quantity is very less, we can drop this



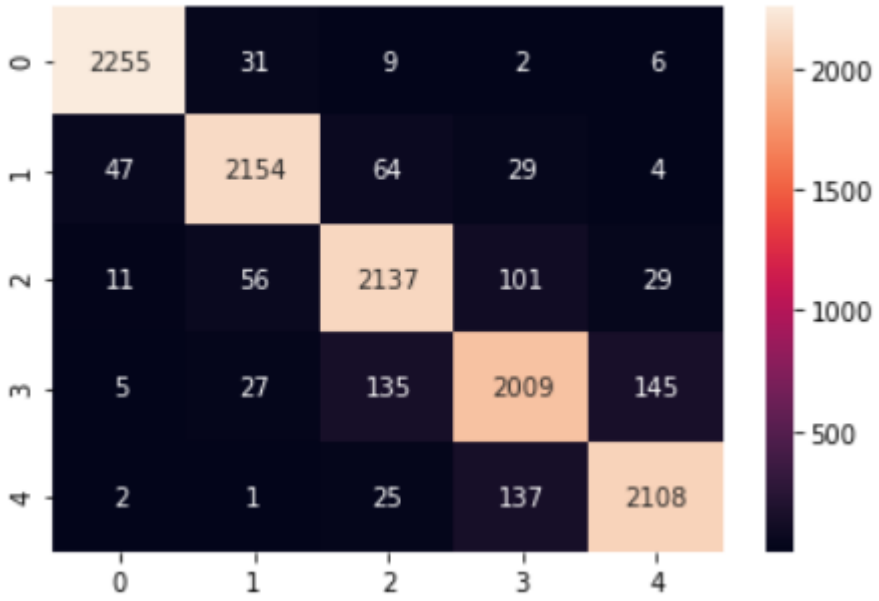
Data Preprocessing

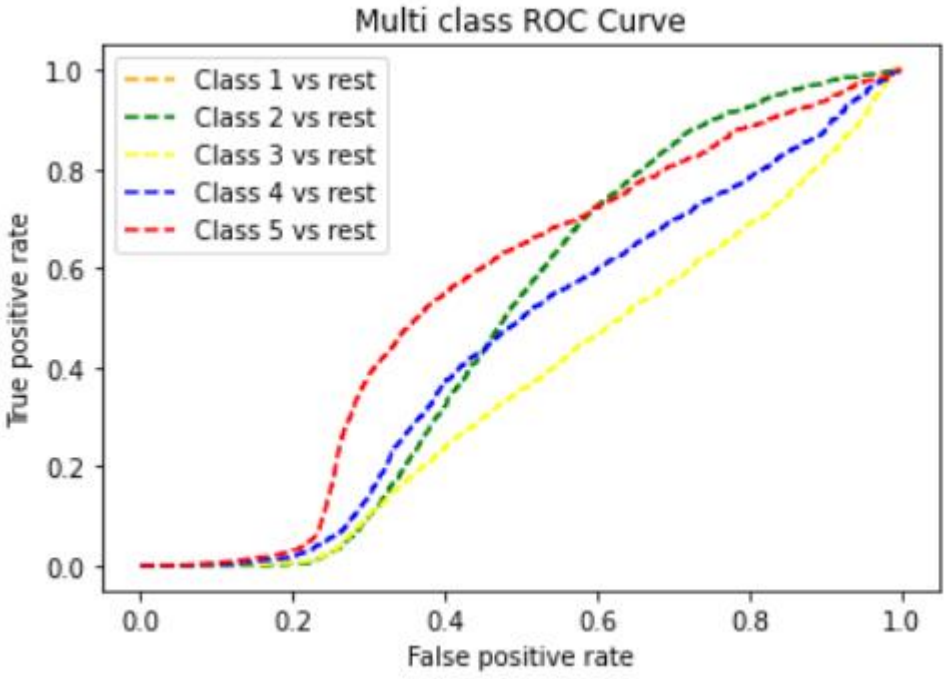
In [43]:
```
corpus=[]

for i in range(len(df)):
    review=re.sub('([a-zA-Z0-9+._-]+@[a-zA-Z0-9._-]+\.[a-zA-Z0-9_-]+)',' ',df['comment'][i])
    review=re.sub('[^a-zA-Z]',' ',df['comment'][i])
    review=review.lower()
    review=review.split()

    review=[lemmit.lemmatize(word) for word in review if word not in set(stopwords.words('english'))]
    review=" ".join(review)
    corpus.append(review)
```

In [44]:
```
corpus
```

Out[44]:
```
['terrible product month bought product convenient use sound bass everything good said water sweat resistance power never worked wearing splashed water still one side stopped working second without warning feel like wasted money wired earphone work longer',
 'absolute rubbish automatically disconnected many time',
 'worst experience ever everything good expect built quality built quality poor another disadvantage use earphone feel pain ear bcz heavy specially right side bud bcz heavy microphone charge socket run bcz due weight stick ear thanks',
 'worst experience ever worst product quality ever stop working one side day',
 'poor bakwas flipkart',
 'poor bad sound quality',
 'poor month used right side speaker working',
 'waste money bad experience sound problem e u call someone person hear u replacement option return acceptable',
 'hated side stopped working day know',
 'meet expectation bad',
 'utterly disappointed fitting earphone good sound ok main issue fitting product worst fitting ever seller accepting return short waste time money please buy',
 'waste money worst product quality never buy product sound quality going day day never bye product',
 'worst experience ever power button working properly used day power button working',
 'meet expectation right side speaker come slowly',
```

Out[76]: <AxesSubplot:>





Conclusion

- Key Findings and Conclusions of the Study

  I used only two websites to scrap data which could be takenas more.

  Data may not be properly preprossed even though applying all the necessary alogithms.

  Knowledge of NLP algorithms is must before working on this dataset.