A PROJECT REPORT

on

SENTIMENT ANALYSIS OF PRODUCTS IN E-COMMERCE PLATFORMS

Submitted to

KIIT Deemed to be University

In Fulfillment of the Requirement for the Award of

BACHELOR'S DEGREE IN
INFORMATION TECHNOLOGY

BY

HARSH MISHRA	2006545
SAYAN CHAKRABORTY	2006085
RISHAV PAL	2006037
ASHISH RAJ	2006538

UNDER THE GUIDANCE OF
SOUMYA RANJAN MISHRA

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA - 751024

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, ODISHA 751024

CERTIFICATE

This is certify that the project entitled

SENTIMENT ANALYSIS OF PRODUCTS IN E-COMMERCE PLATFORMS

submitted by

HARSH MISHRA	2006545
SAYAN CHAKRABORTY	2006085
RISHAV PAL	2006037
ASHISH RAJ	2006538

is a record of bonafide work carried out by them, in the fulfilment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering OR Information Technology) at KIIT Deemed to be university, Bhubaneswar. This work is done during year 2022-2023, under our guidance.

Date: / /

(Guide Name)
Project Guide

Acknowledgement

We are profoundly grateful to **SOUMYA RANJAN MISHRA** sir of **Affiliation** for his expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion.

HARSH MISHRA
RISHAV PAL
SAYAN CHAKRABORTY
ASHISH RAJ

ABSTRACT

In recent years, with the rapid development of Internet technology, online shopping has become a common way for users to shop and consume. Sentiment analysis of a large number of user reviews on e-commerce platforms can effectively improve user satisfaction. The world we see today is becoming more and more digitized. In this digitized world, e-commerce is gaining momentum by making products available to customers without leaving their homes. As people today rely on online products, the importance of reviews is increasing. Before a customer can choose a product, they have to go through thousands of reviews to understand the product. But in today's booming age of machine learning, sifting through thousands of reviews would be much easier if a model is used to polarize those reviews and learn from them.

In the proposed work, 4 datasets (Flip-kart, Myntra, Amazon, Apple) were classified into positive, neutral and negative sentiments using Sentiment Analysis. Of the various classification models, Naïve Bayes was used to classify the reviews.

Contents

Introduction

As the commercial site of the world is almost fully undergone in online platform people is trading products through different e commerce website. And for that reason reviewing products before buying is also a common scenario. Also now a day, customers are more inclined towards the reviews to buy a product. So analyzing the data from those customer reviews to make the data more dynamic is an essential field nowadays. In this age of increasing machine learning based algorithms reading thousands of reviews to understand a product is rather time consuming where we can polarize a review on particular category to understand its popularity among the buyers all over the world.

The objective of this paper is to categorize the positive and negative feedback of the customers over different products and build a supervised learning model to polarize large amount of reviews. A study on amazon last year revealed over 88% of online shoppers trust reviews as much as personal recommendations. Any online item with large amount of positive reviews provides a powerful comment of the legitimacy of the item. Conversely, books, or any other online item, without reviews puts potential prospects in a state of distrust. Quite simply, more reviews look more convincing.

Chapter 2

Methodology

2.1 Data collection

In our analysis, we utilized four distinct datasets with the aim of improving user experience. The first dataset was sourced from Myntra and included a total of 2782 tuples. This dataset had attributes such as ASIN, name, date, rating, and review. The rating attribute referred to the rating given by the user for a product, while the review attribute contained the user's written review of the product.The second dataset was sourced from Flipkart and contained 2304 tuples. The dataset had attributes such as product name, review, and rating. Similar to the Myntra dataset, the rating attribute referred to the rating given by the user for a product, while the review attribute contained the user's written review of the product.The third dataset was sourced from Apple and included a total of 9713 tuples. This dataset had attributes such as ratings, comments, and reviews. The ratings attribute referred to the rating given by the user for an app, while the comments attribute contained the user's written feedback on the app.Finally, we utilized a dataset sourced from Amazon, which contained 1465 tuples. This dataset had attributes such as product_name, category, Rating, about_product, and user_id. The category attribute referred to the category of the product, while the about_product attribute contained information about the product.

To begin our analysis, we imported all the CSV files into our system and proceeded to compare the different datasets to identify areas where we could enhance the user experience.

The dataset are as follows:-

1) Myntra

```
df
```

	asin	name	date	rating	review
0	B07W7CTLD1	Mamaearth-Onion-Growth-Control-Redensyl	9/6/2019	1	I bought this hair oil after viewing so many g...
1	B07W7CTLD1	Mamaearth-Onion-Growth-Control-Redensyl	8/14/2019	5	Used This Mama Earth Newly Launched Onion Oil ...
2	B07W7CTLD1	Mamaearth-Onion-Growth-Control-Redensyl	10/19/2019	1	So bad product...My hair falling increase too ...
3	B07W7CTLD1	Mamaearth-Onion-Growth-Control-Redensyl	9/16/2019	1	Product just smells similar to navarathna hair...
4	B07W7CTLD1	Mamaearth-Onion-Growth-Control-Redensyl	8/18/2019	5	I have been trying different onion oil for my ...
...	...	...	...	...	...
2777	B07MVHJ6CH	Mysore-Sandal-Soaps-Pack-Bars	3/1/2020	5	Long lasting freshness throughout the day.
2778	B07MVHJ6CH	Mysore-Sandal-Soaps-Pack-Bars	10/24/2019	5	My preferred soap
2779	B07MVHJ6CH	Mysore-Sandal-Soaps-Pack-Bars	10/3/2020	2	Fantastic
2780	B07MVHJ6CH	Mysore-Sandal-Soaps-Pack-Bars	6/21/2019	4	Super Product
2781	B07MVHJ6CH	Mysore-Sandal-Soaps-Pack-Bars	7/3/2020	5	Best soothing, cooling fragrance for hot summe...

2782 rows × 5 columns

2) Flipkart

```
df.head()
```

	Unnamed: 0	Product_name	Review	Rating
0	0	Lenovo Ideapad Gaming 3 Ryzen 5 Hexa Core 5600...	Best under 60k Great performance! got it for a...	5
1	1	Lenovo Ideapad Gaming 3 Ryzen 5 Hexa Core 5600...	Good perfomence...	5
2	2	Lenovo Ideapad Gaming 3 Ryzen 5 Hexa Core 5600...	Great performance but usually it has also that...	5
3	3	DELL Inspiron Athlon Dual Core 3050U - (4 GB/2...	My wife is so happy and best product 👌😍	5
4	4	DELL Inspiron Athlon Dual Core 3050U - (4 GB/2...	Light weight laptop with new amazing features,...	5

```
df.tail()
```

	Unnamed: 0	Product_name	Review	Rating
2299	2299	MSI 27 inch Full HD IPS Panel Monitor (PRO MP2...	Great display, accurate colours at this price ...	5
2300	2300	MSI 27 inch Full HD IPS Panel Monitor (PRO MP2...	Superb monitor first brought 1 used for 2 mont...	5
2301	2301	MSI 27 inch Full HD IPS Panel Monitor (PRO MP2...	Awesome	5
2302	2302	MSI 27 inch Full HD IPS Panel Monitor (PRO MP2...	Only one issue with adapter	5
2303	2303	MSI 27 inch Full HD IPS Panel Monitor (PRO MP2...	Worth the money u spend for this monitor Great...	5

3) Apple

```
df
```

	Ratings	Comment	Reviews
0	5	Super!	Great camera for pics and videos Battery life ...
1	5	Must buy!	Great device. Let me tell the Pros..1. Superb ...
2	5	Great product	Who all loves older size i.e., 4.7 inch type s...
3	5	Simply awesome	This iPhone SE is the best phone ever you get....
4	5	Classy product	This is my second iphone after iphone 4s. I've...
...	...	...	...
9708	5	Terrific purchase	Absolutely brilliantREAD MORE
9709	5	Classy product	Superb phone. This is my 4th iPhone, I feel SE...
9710	5	Awesome	very niceREAD MORE
9711	5	Super!	Loving it as of now. Good Product .READ MORE
9712	5	Terrific purchase	NiceElegant ElectricREAD MORE

9713 rows × 3 columns

4) Amazon

```
df.head()
```

	product_name	category	Rating	about_product	user_id	
0	Wayona Nylon Braided USB to Lightning Fast Cha...	Computers&Accessories|Accessories&Peripherals|...	4.2	High Compatibility : Compatible With iPhone 12...	AG3D6O4STAQKAY2UVGEUV46KN35Q,AHMY5CWJMMK5BJRBB...	R3
1	Ambrane Unbreakable 60W / 3A Fast Charging 1.5...	Computers&Accessories|Accessories&Peripherals|...	4	Compatible with all Type C enabled devices, be...	AECPFYFQVRUWC3KGNLJIOREFP5LQ,AGYYVPDD7YG7FYNBX...	RG
2	Sounce Fast Phone Charging Cable & Data Sync U...	Computers&Accessories|Accessories&Peripherals|...	3.9	【 Fast Charger& Data Sync】-With built-in safet...	AGU3BBQ2V2DDAMOAKGFAWDDQ6QHA,AESFLDV2PT363T2AQ...	
3	boAt Deuce USB 300 2 in 1 Type-C & Micro USB S...	Computers&Accessories|Accessories&Peripherals|...	4.2	The boAt Deuce USB 300 2 in 1 cable is compati...	AEWAZDZZJLQUYVOVGBEUKSLXHQ5A,AG5HTSFRRE6NL3M5S...	R
4	Portronics Konnect L 1.2M Fast Charging 3A 8 P...	Computers&Accessories|Accessories&Peripherals|...	4.2	[CHARGE & SYNC FUNCTION]- This cable comes wit...	AE3Q6KSUK5P75D5HFYHCRAOLODSA,AFUGIFH5ZAFXRDSZH...	F
1464	Sandwich Maker (...	Home&Kitchen|Kitchen&HomeAppliances|SmallKitch...	4.3	ac 23V – 5Hz;1 W...	AFGW5PT3R6ZAVQR4Y5MWVAKBZAYA,AG7QNJ2SCS5VS5VY	

2.2 Data Pre-processing

One of the primary challenges with the datasets we used is the presence of missing values across various attributes. These missing values can significantly reduce the accuracy of the Machine Learning models and hinder their overall efficiency.To address this issue, we implemented a method to replace the missing values with the mean value of the respective column. This approach involves substituting the missing value with the mean value of the neighboring values. By doing so, we obtain an approximate optimal value for the missing attribute.We achieved this by first calculating the mean value of the column and then substituting the missing value with the calculated mean value. This helped us ensure that we have a complete dataset with minimal missing values, thereby enabling us to enhance the accuracy and efficiency of our Machine Learning models.

For **Myntra dataset** we have :

```
df.duplicated().sum()
```

```
1379
```

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

```
asin      0
name      0
date      0
rating    0
review    0
dtype: int64
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2782 entries, 0 to 2781
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   asin    2782 non-null   object
 1   name    2782 non-null   object
 2   date    2782 non-null   object
 3   rating  2782 non-null   int64 
 4   review  2782 non-null   object
dtypes: int64(1), object(4)
memory usage: 108.8+ KB
```

For **Flipkart dataset** we have:-

```
df.duplicated().sum()
```

```
0
```

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

```
Unnamed: 0      0
Product_name    0
Review          0
Rating          0
dtype: int64
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2304 entries, 0 to 2303
Data columns (total 4 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   Unnamed: 0     2304 non-null   int64 
 1   Product_name   2304 non-null   object
 2   Review         2304 non-null   object
 3   Rating         2304 non-null   int64 
dtypes: int64(2), object(2)
memory usage: 72.1+ KB
```

For **Apple datset** we have:-

```
df.duplicated().sum()
```

```
649
```

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

```
Ratings    0
Comment    0
Reviews    0
dtype: int64
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9713 entries, 0 to 9712
Data columns (total 3 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   Ratings  9713 non-null   int64 
 1   Comment  9713 non-null   object
 2   Reviews  9713 non-null   object
dtypes: int64(1), object(2)
memory usage: 227.8+ KB
```

For **Amazon dataset** we have:-

```
df.duplicated().sum()
```

```
100
```

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

```
product_name      0
category          0
Rating            0
about_product     0
user_id           0
review_id         0
review_title      0
Review            0
dtype: int64
```

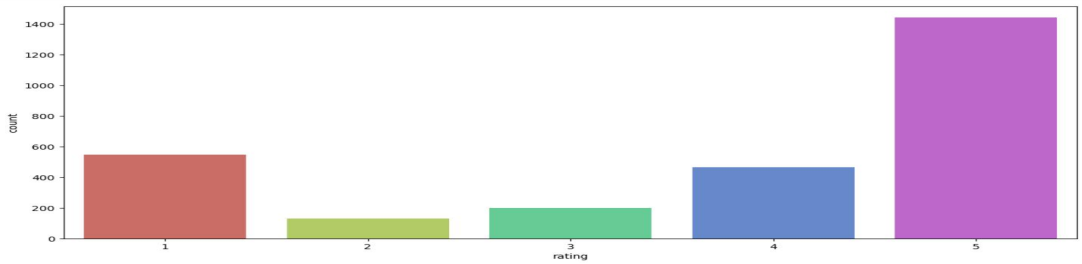
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1465 entries, 0 to 1464
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   product_name    1465 non-null   object
 1   category        1465 non-null   object
 2   Rating          1465 non-null   object
 3   about_product   1465 non-null   object
 4   user_id         1465 non-null   object
 5   review_id       1465 non-null   object
 6   review_title    1465 non-null   object
 7   Review          1465 non-null   object
dtypes: object(8)
memory usage: 91.7+ KB
```

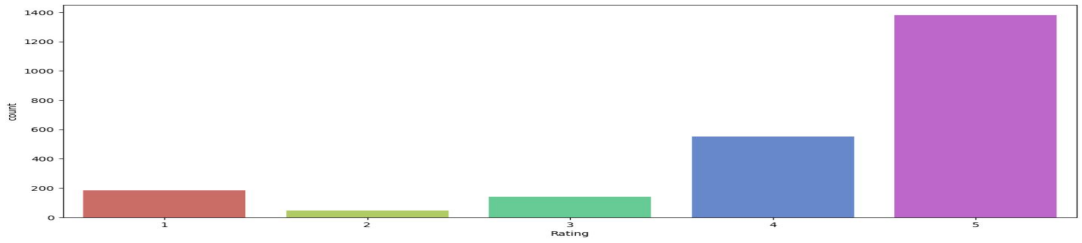
2.3 Data visualization

Graph between Rating and count:-

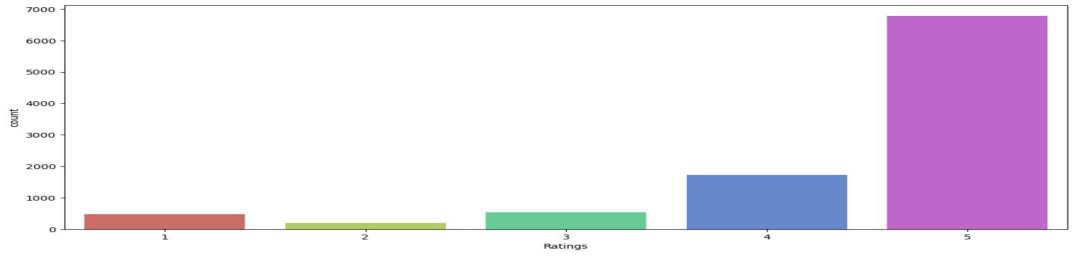
1) Myntra



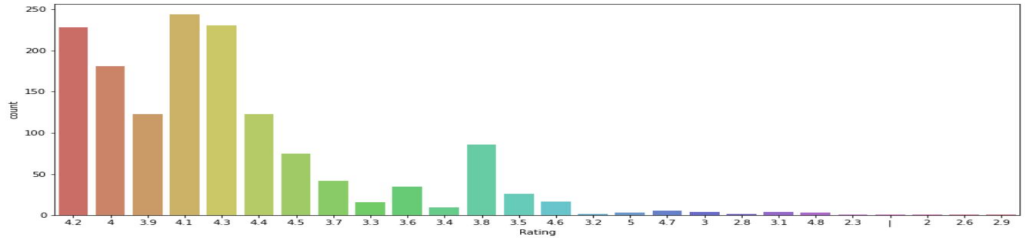
2) Flipkart



3) Apple



4) Amazon

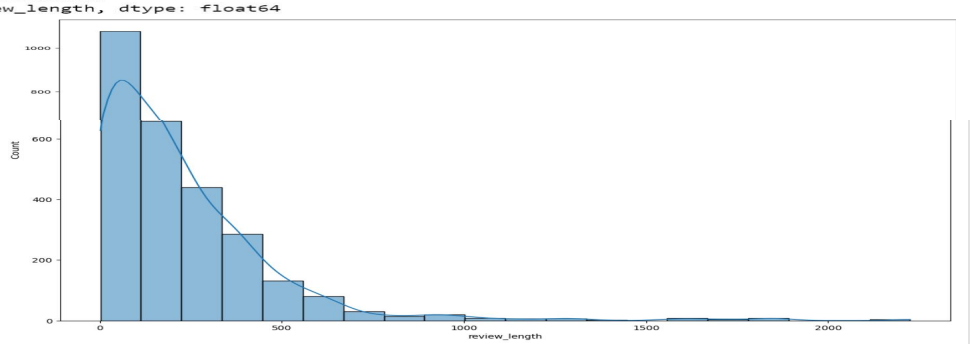


Count and review length

1) Myntra

description

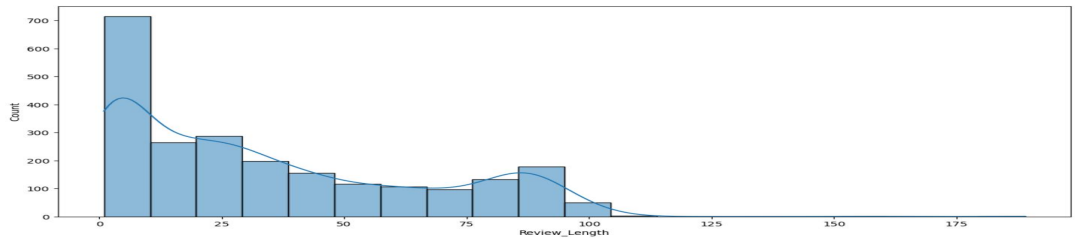
```
count     2782.000000
mean       224.400072
std        249.820358
min          2.000000
25%         57.000000
50%        159.000000
75%        310.750000
max       2225.000000
Name: review_length, dtype: float64
```



2) Flipkart

description

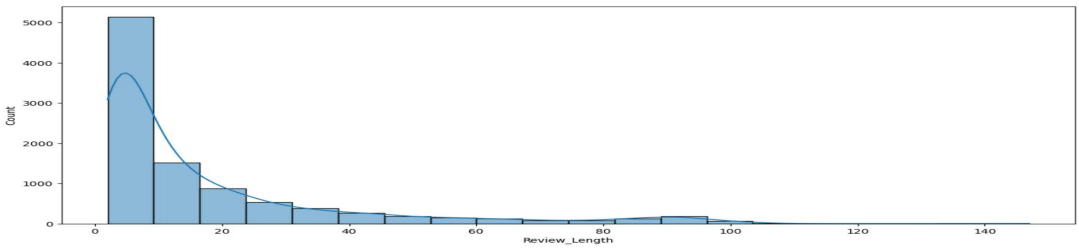
```
count    2304.000000
mean       34.242622
std        30.535511
min         1.000000
25%         6.000000
50%        26.000000
75%        56.000000
max       189.000000
Name: Review_Length, dtype: float64
```



3) Apple

description

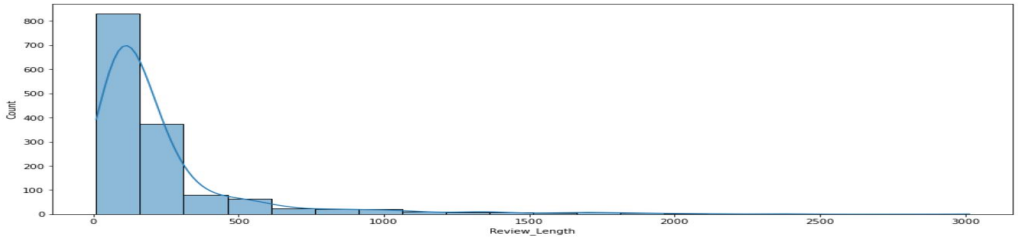
```
count    9713.000000
mean       17.542469
std        21.680915
min         2.000000
25%         4.000000
50%         9.000000
75%        21.000000
max       147.000000
Name: Review_Length, dtype: float64
```



4) Amazon

description

```
count    1465.000000
mean      241.630034
std       317.550129
min        11.000000
25%        81.000000
50%       140.000000
75%       243.000000
max      3014.000000
Name: Review_Length, dtype: float64
```

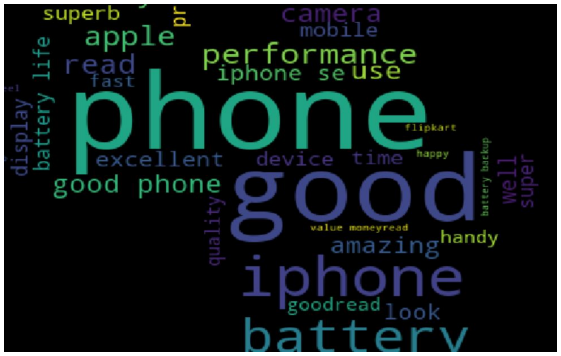


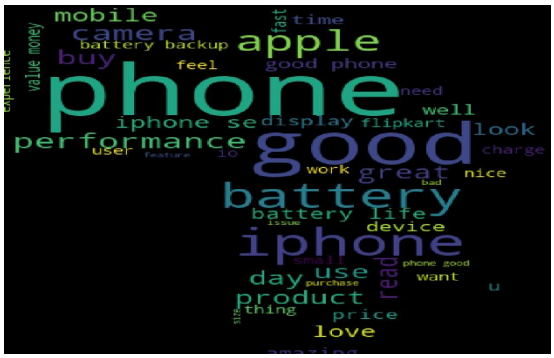
Top words for 5 rating and review length greater than 11

1) Myntra

2) Flipkart

3) Apple



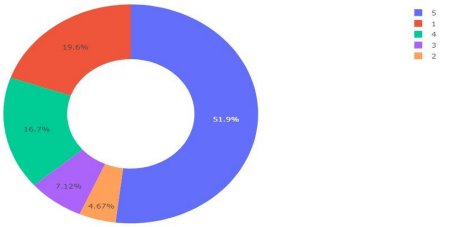


4) Amazon

Pie chart between ratings and count

1) Myntra

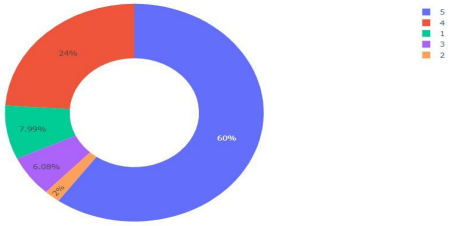
```
In [71]: figure = px.pie(df_new, values=quantity, names=numbers,hole = 0.5)
         figure.show()
```



2) Flipkart

```
In [77]: ratings = df['Rating'].value_counts()
         numbers = ratings.index
         quantity = ratings.values
```

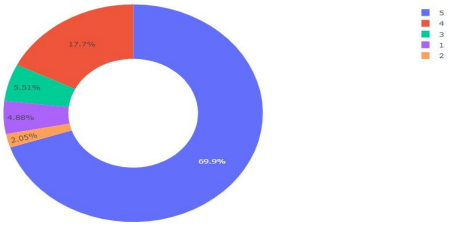
```
In [78]: figure = px.pie(df_new, values=quantity, names=numbers,hole = 0.5)
         figure.show()
```



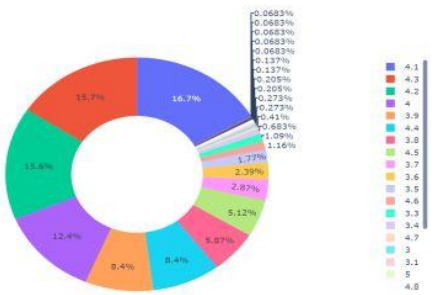
3) Apple

```
In [41]: ratings = df['Ratings'].value_counts()
         numbers = ratings.index
         quantity = ratings.values
```

```
In [42]: figure = px.pie(df_new, values=quantity, names=numbers,hole = 0.5)
         figure.show()
```



4) Amazon



Chapter 3
RESULT

1) Myntra

```
x = sum(df_new["Positive"])
y = sum(df_new["Negative"])
z = sum(df_new["Neutral"])
def sentiment_score(a, b, c):
 if (a>b) and (a>c):
    print("Positive 😊 ")
 elif (b>a) and (b>c):
    print("Negative 😠 ")
 else:
    print("Neutral 😐 ")
sentiment_score(x, y, z)
```

```
Neutral 😐
```

```
print("Positive: ", x)
print("Negative: ", y)
print("Neutral: ", z)
```

```
Positive:  708.7869999999996
Negative:  155.26000000000016
Neutral:  1917.9730000000015
```

2) Flipkart

```
x = sum(df_new["Positive"])
y = sum(df_new["Negative"])
z = sum(df_new["Neutral"])
def sentiment_score(a, b, c):
    if (a>b) and (a>c):
        print("Positive 😊 ")
    elif (b>a) and (b>c):
        print("Negative 😠 ")
    else:
        print("Neutral 😐 ")
sentiment_score(x, y, z)
```

```
Positive 😊
```

```
print("Positive: ", x)
print("Negative: ", y)
print("Neutral: ", z)
```

```
Positive:  1118.6070000000016
Negative:  106.38300000000005
Neutral:  1079.0409999999997
```

3) Apple

```
x = sum(df_new["Positive"])
y = sum(df_new["Negative"])
z = sum(df_new["Neutral"])
def sentiment_score(a, b, c):
    if (a>b) and (a>c):
        print("Positive 😊 ")
    elif (b>a) and (b>c):
        print("Negative 😠 ")
    else:
        print("Neutral 😐 ")
sentiment_score(x, y, z)
```

```
Neutral 😐
```

```
print("Positive: ", x)
print("Negative: ", y)
print("Neutral: ", z)
```

```
Positive:  3506.43200000009
Negative:  327.47399999999965
Neutral:  5878.109999999949
```

4) Amazon

```
x = sum(df_new["Positive"])
y = sum(df_new["Negative"])
z = sum(df_new["Neutral"])
def sentiment_score(a, b, c):
    if (a>b) and (a>c):
        print("Positive 😊 ")
    elif (b>a) and (b>c):
        print("Negative 😠 ")
    else:
        print("Neutral 😐 ")
sentiment_score(x, y, z)
```

```
Neutral 😐
```

```
print("Positive: ", x)
print("Negative: ", y)
print("Neutral: ", z)
```

```
Positive:  430.30600000000055
Negative:  83.35400000000011
Neutral:  951.3630000000026
```

Conclusion

In this research we proposed a supervised learning model to polarize a large amount of product review dataset which was unlabeled. We proposed our model which is a supervised learning method and used a mix of 2 kinds of feature extractor approach. We described the basic theory behind the model, approaches we used in our research and the performance measure for the conducted experiment over quite a large data. We also compared our result with some of the similar works regarding product review. We also went through different kinds of research papers regarding sentiment analysis over a text based dataset. Accuracy of the dataset using Naive Bayes:

1. Myntra- 76.14%
2. Flipkart- 62.30%
3. Apple- 70.19%
4. Amazon- 26.43%

Future Scope

Some future works which can be included to improve the model and also to make it more effective in practical cases. Our future works include applying PCA (Principal Component Analysis) in active learning process to fully automate data labeling process with less assistance from the oracle. The model can be incorporate with programs that can interact with customer seeking a score of a particular product. As we used a large scale dataset we can apply the model on local market sites to get better accuracy and usability. And lastly we will try to continue this research until we generalize this model to all kinds of text based reviews and comments.