CSE508 Information Retrieval
Assignment-3

Megha 2021337

1. **Product chosen : Speaker**

‘Speaker.csv’ : Product details

‘User_review.csv’ : Customer reviews

3.**Total Number of items :** 31646 rows

```
speaker_df = df[df['title'].str.contains('speaker', case=False, na=False)]

print("Rows containing the keyword 'speaker' in the 'title' column:")
speaker_df
speaker_df.dropna(inplace=True)
speaker_df = speaker_df.drop_duplicates(subset='asin', keep='first')
speaker_df
```

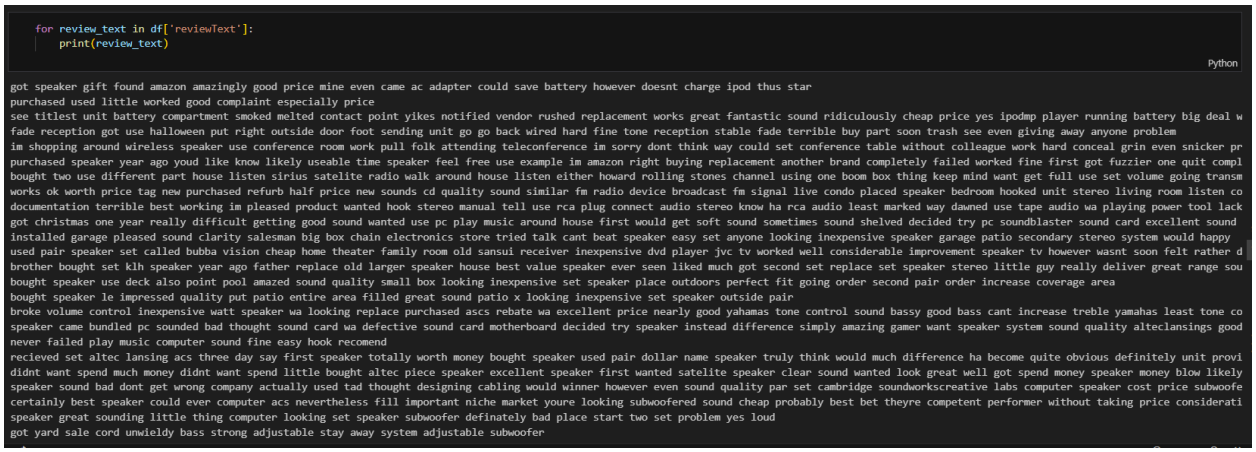
After Removing duplicate rows and handling missing values :

30195 rows

4.The Descriptive Statistics of the product :

```
Number of Reviews: 306136
Average Rating Score: 4.273820098493854
Number of Unique Products: 30195
Number of Good Ratings: 273451
Number of Bad Ratings: 32761
Number of Reviews corresponding to each Rating:
overall
1.0     18254
2.0     14507
3.0     24620
4.0     56588
5.0    192243
Name: count, dtype: int64
```

5.Preprocessing User review :



```
for review_text in df['reviewText']:
    print(review_text)
```

Python

```
got speaker gift found amazon amazingly good price mine even came ac adapter could save battery however doesnt charge ipod thus star
purchased used little worked good complaint especially price
see titlest unit battery compartment smoked melted contact point yikes notified vendor rushed replacement works great fantastic sound ridiculously cheap price yes ipodmp player running battery big deal w
fade reception got use halloween put right outside door foot sending unit go go back wired hard fine tone reception stable fade terrible buy part soon trash see even giving away anyone problem
im shopping around wireless speaker use conference room work pull folk attending teleconference im sorry dont think way could set conference table without colleague work hard conceal grin even snicker pr
purchased speaker year ago youd like know likely useable time speaker feel free use example im amazon right buying replacement another brand completely failed worked fine first got fuzzier one quit compl
bought two use different part house listen sirius satelite radio walk around house listen either howard rolling stones channel using one boom box thing keep mind want get full use set volume going transm
works ok worth price tag new purchased refurb half price new sounds cd quality sound similar fm radio device broadcast fm signal live condo placed speaker bedroom hooked unit stereo living room listen co
documentation terrible best working im pleased product wanted hook stereo manual tell use rca plug connect audio stereo know ha rca audio least marked way dawned use tape audio wa playing power tool lack
got christmas one year really difficult getting good sound wanted use pc play music around house first would get soft sound sometimes sound shelved decided try pc soundblaster sound card excellent sound
installed garage pleased sound clarity salesman big box chain electronics store tried talk cant beat speaker easy set anyone looking inexpensive speaker garage patio secondary stereo system would happy
used pair speaker set called bubba vision cheap home theater family room old sansui receiver inexpensive dvd player jvc tv worked well considerable improvement speaker tv however wasnt soon felt rather d
brother bought set klh speaker year ago father replace old larger speaker house best value speaker ever seen liked much got second set replace set speaker stereo little guy really deliver great range sou
bought speaker use deck also point pool amazed sound quality small box looking inexpensive set speaker place outdoors perfect fit going order second pair order increase coverage area
bought speaker le impressed quality put patio entire area filled great sound patio x looking inexpensive set speaker outside pair
broke volume control inexpensive watt speaker wa looking replace purchased ascs rebate wa excellent price nearly good yahamas tone control sound bassy good bass cant increase treble yamahas least tone co
speaker came bundled pc sounded bad thought sound card wa defective sound card motherboard decided try speaker instead difference simply amazing gamer want speaker system sound quality alteclansings good
never failed play music computer sound fine easy hook recomend
recieved set altec lansing acs three day say first speaker totally worth money bought speaker used pair dollar name speaker truly think would much difference ha become quite obvious definitely unit provi
didnt want spend much money didnt want spend little bought altec piece speaker excellent speaker first wanted satelite speaker clear sound wanted look great well got spend money speaker money blow likely
speaker sound bad dont get wrong company actually used tad thought designing cabling would winner however even sound quality par set cambridge soundworkscreative labs computer speaker cost price subwoofe
certainly best speaker could ever computer acs nevertheless fill important niche market youre looking subwoofered sound cheap probably best bet theyre competent performer without taking price considerati
speaker great sounding little thing computer looking set speaker subwoofer definately bad place start two set problem yes loud
got yard sale cord unwieldy bass strong adjustable stay away system adjustable subwoofer
```

6. EDA :

Top 20 most reviewed brands:

Brand	Number of Reviews
1. Logitech	18519
2. Pyle	9601
3. Bose	9169
4. JBL	8830
5. AmazonBasics	8644
6. Sony	6402
7. Cambridge Soundworks	6293
8. Polk Audio	6189
9. Anker	4279
10. Altec Lansing	3961
11. Monoprice	3492
12. Metra	3258
13. Loopilops	3073
14. Pioneer	3068
15. iHome	3006
16. Yamaha Audio	3002
17. AYL	2871
18. Klipsch	2854
19. BOSS Audio Systems	2740
20. Philips	2695

Top 20 least reviewed brands:

Brand	Number of Reviews

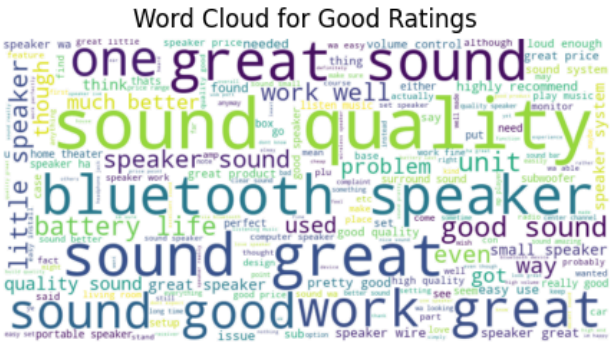
1. Exlight 5
2. Sunvito 5
3. Best Kits 5
4. Pagreberya 5
5. Ukonnect 5
6. Aike 5
7. iwerkz 5
8. Pomelo Best 5
9. Foho 5
10. ACCEX 5
11. Sinool 5
12. Geekria 5
13. Eathtek 5
14. Fatboy Sounds 5
15. SeaSum 5
16. BSWHW 5
17. Teknub 5
18. Armor 5
19. Fred & Friends 4
20. ZOpid 3

Most reviewed brand:

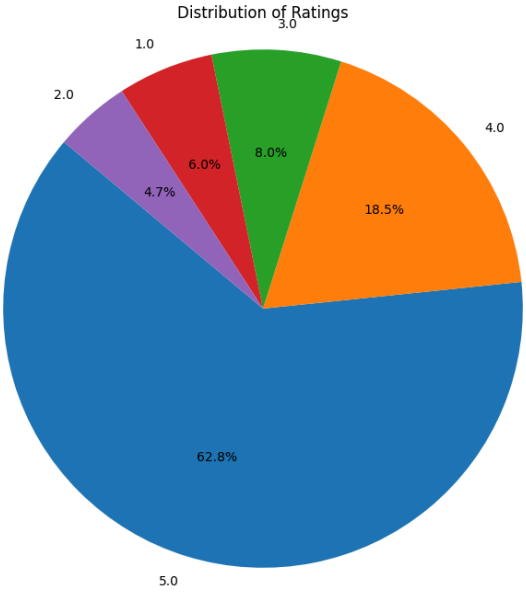
```
Most positively reviewed product: B010OYASRG
Average rating: 4.6769480519480515
Brand associated  Cambridge Soundworks
```

overall	1.0	2.0	3.0	4.0	5.0
year					
2015	13	13	24	76	294
2016	83	67	130	337	1455
2017	90	76	155	312	1385
2018	38	44	59	131	570

Word Cloud



Distribution of Rating Vs Number of Reviews in %.




```
Year with the highest number of verified customers: 2016
Number of verified customers in that year: 53296
```

7.TF-IDF matrix creation

```
(0, 42255)    0.1393889793801603
(0, 10946)    0.1127921109101254
(0, 57659)    0.08611346004085624
(0, 28212)    0.1465137305014746
(0, 53917)    0.20493699895220654
(0, 39581)    0.105423787873217
(0, 38774)    0.12736602833481223
(0, 57881)    0.06742819678653644
(0, 52164)    0.17754205301638312
(0, 12452)    0.29567829189226796
(0, 58406)    0.09988817157836341
(0, 30685)    0.2117972055818393
(0, 28630)    0.1371624006342914
(0, 21950)    0.08676776306968237
(0, 28035)    0.11645532052119445
(0, 3180)     0.3146038696678656
(0, 10576)    0.12253555750249495
(0, 38948)    0.11623289126540248
(0, 41921)    0.34577020510735224
(0, 56735)    0.15723527420098002
(0, 52443)    0.14165652360361192
(0, 30576)    0.15993346514837095
(0, 50242)    0.25090598622693133
(0, 23565)    0.15541748765134664
(0, 58106)    0.12376679216329935
```

```
Feature: wirelessfree, IDF: 11.797910927025205
Feature: wirelessgreat, IDF: 11.797910927025205
Feature: wirelessit, IDF: 11.797910927025205
Feature: wirelessly, IDF: 6.965605168453367
Feature: wirelesslyinductively, IDF: 11.797910927025205
Feature: wirelessmicro, IDF: 11.797910927025205
Feature: wirelessmy, IDF: 11.797910927025205
Feature: wirelessn, IDF: 11.10476374646526
```

8.Dividing Rating Class

Old data set

```
Class
Good       248432
Bad         32737
Average     24599
Name: count, dtype: int64 Class
Good       81.248528
Bad        10.706483
Average     8.044988
Name: count, dtype: float64
```

Modified data set

```
Class
Good       40500
Bad        32737
Average    24599
Name: count, dtype: int64 Class
Good       41.395805
Bad        33.461098
Average    25.143097
Name: count, dtype: float64
```

The data is unevenly distributed. This has caused poor performance for the classes labeled as ‘Average’ and ‘Bad’.

9.Dataset is divided into train.csv and test.csv

```
from sklearn.model_selection import train_test_split

X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.25,random_state=

train_df=pd.DataFrame({'reviewText':X_train,'Class':y_train})
test_df=pd.DataFrame({'reviewText':X_test,'Class':y_test})

train_df.to_csv('train.csv',index=False)
test_df.to_csv('test.csv',index=False)
```

10.Model training on our dataset

Naive Bayes

	precision	recall	f1-score	support
Average	0.70	0.11	0.20	6221
Bad	0.68	0.77	0.72	8096
Good	0.65	0.92	0.76	10142
accuracy			0.67	24459
macro avg	0.68	0.60	0.56	24459
weighted avg	0.68	0.67	0.61	24459

Linear SVC

	precision	recall	f1-score	support
Average	0.57	0.43	0.49	6221
Bad	0.71	0.76	0.74	8096
Good	0.77	0.84	0.81	10142
accuracy			0.71	24459
macro avg	0.69	0.68	0.68	24459
weighted avg	0.70	0.71	0.70	24459

Random Forest

	precision	recall	f1-score	support
Average	0.67	0.31	0.42	6221
Bad	0.69	0.81	0.75	8096
Good	0.73	0.87	0.80	10142
accuracy			0.71	24459
macro avg	0.70	0.66	0.65	24459
weighted avg	0.70	0.71	0.68	24459

Gradient boosting

	precision	recall	f1-score	support
Average	0.58	0.31	0.40	6221
Bad	0.63	0.75	0.68	8096
Good	0.70	0.79	0.74	10142
accuracy			0.65	24459
macro avg	0.63	0.62	0.61	24459
weighted avg	0.64	0.65	0.64	24459

MLP Classifier

Classification Report for MLPClassifier:

	precision	recall	f1-score	support
Average	0.38	0.30	0.33	6055
Bad	0.68	0.63	0.66	8238
Good	0.92	0.95	0.94	62149
accuracy			0.86	76442
macro avg	0.66	0.63	0.64	76442
weighted avg	0.85	0.86	0.86	76442

11.Collaborative filtering :

1.User_item_matrix

2.Normalized rating (0 to 1)