Product Review Rating Predication

BY- RAJ SHARMA

INTRODUCTION

- Research has shown that consumer online product ratings reflect both the customers' experience with the product and the influence of others' ratings.
- The opinion information is very useful for users and customers. Businesses can also use the opinion information to design better strategies for production and marketing.
- A recent survey (Hinckley, 2015) revealed that 67.7% of consumers are effectively influenced by online reviews when making their purchase decisions.
- More precisely, 54.7% recognized that these reviews were either fairly, very or absolutely important in their purchase decision making.
- Searching and comparing text reviews can be frustrating for users as they feel submerged with information.

INTRODUCTION

- The star-rating, i.e., stars from 1 to 5 on online platform, rather than its text content gives a quick overview of the product quality.
- The overall star ratings of the product reviews may not capture the exact polarity of the sentiments.
- For instance, a user may rate a product as good and assign a 5-star score while another user may write the same comment and give only 3 stars.
- *The question that arises is how to successfully predict a user's numerical rating from its review text content*. One solution is to rely on supervised machine learning techniques such as text classification which allows to automatically classify a document into a fixed set of classes after being trained over past annotated data.

PROBLEM STATEMENT

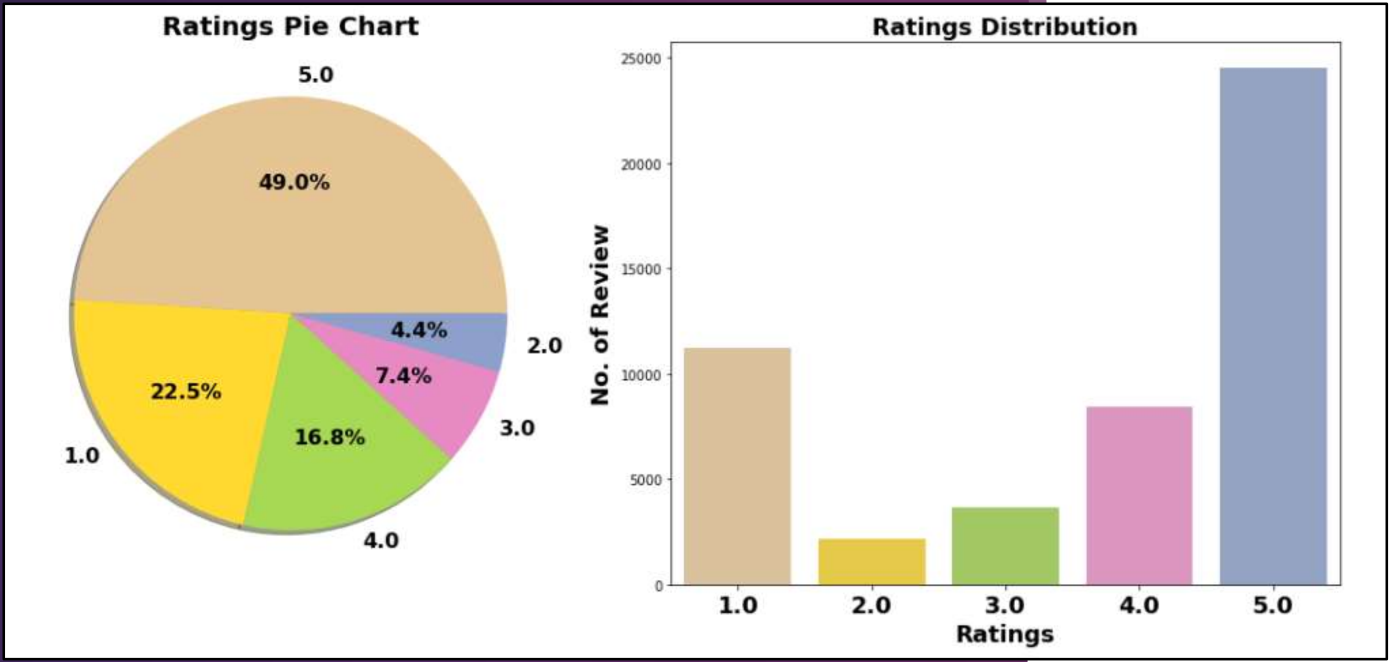
We have a client who has a website where people write different reviews for technical products. Now they are adding a new feature to their website i.e. The reviewer will have to add stars(rating) as well with the review. The rating is out 5 stars and it only has 5 options available 1 star, 2 stars, 3 stars, 4 stars, 5 stars. Now they want to predict ratings for the reviews which were written in the past and they don't have rating.

So we, we have to build an application which can predict the rating by seeing the review.

Web Scraping Details

- Web Scraping done using selenium web driver.
- Data for different product like smartphones, laptops, routers is scraped.
- Data scraped from amazon.in & Flipkart.com
- Around 50000 product reviews are scrap for this project.

Exploration of Target Variable Ratings



Comment:

1. Around 49% customer given 5- star rating followed by 22.5% customer given lowest 1-star rating.

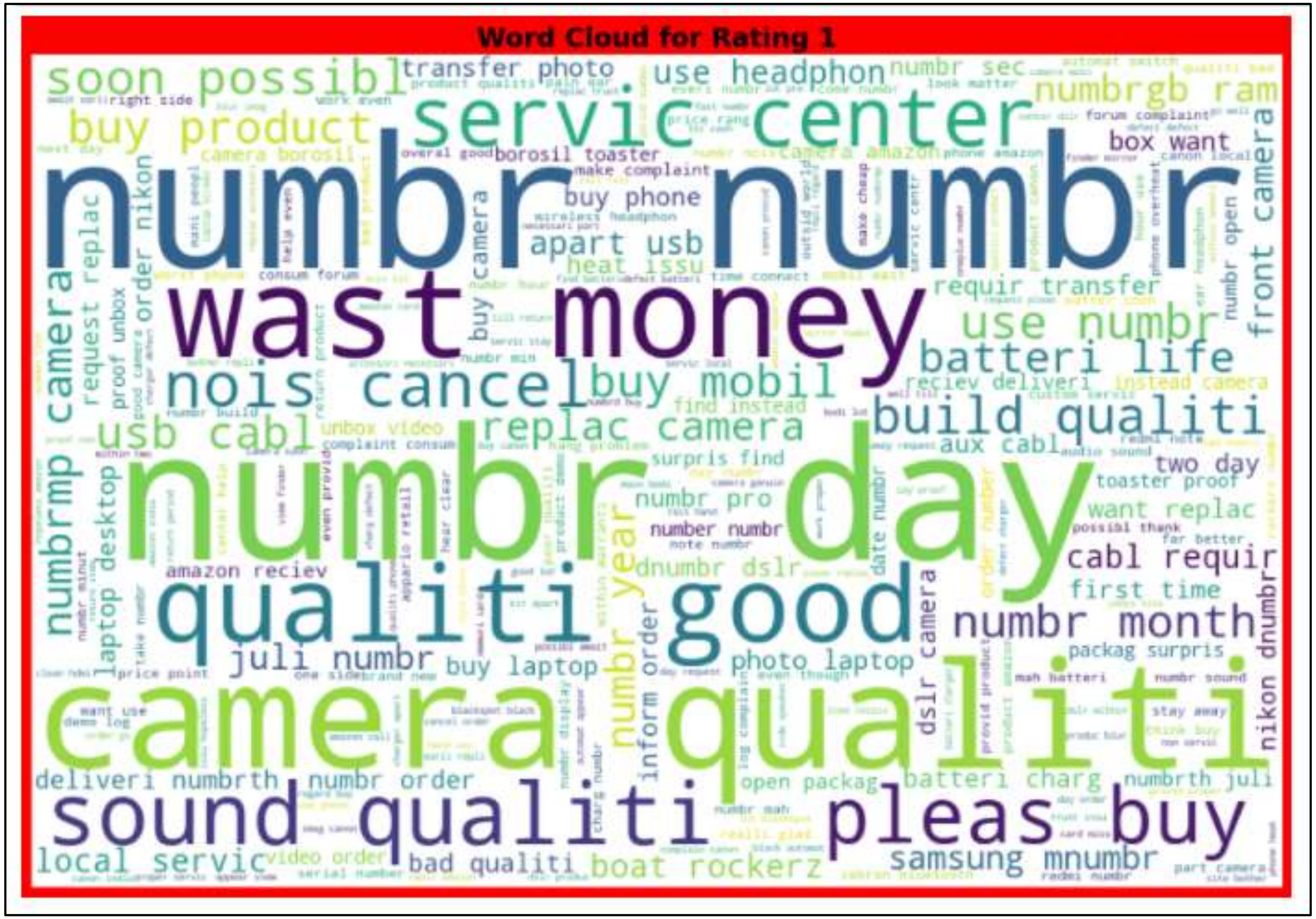
2. Average Rating is 3.65.

Data Pre Processing

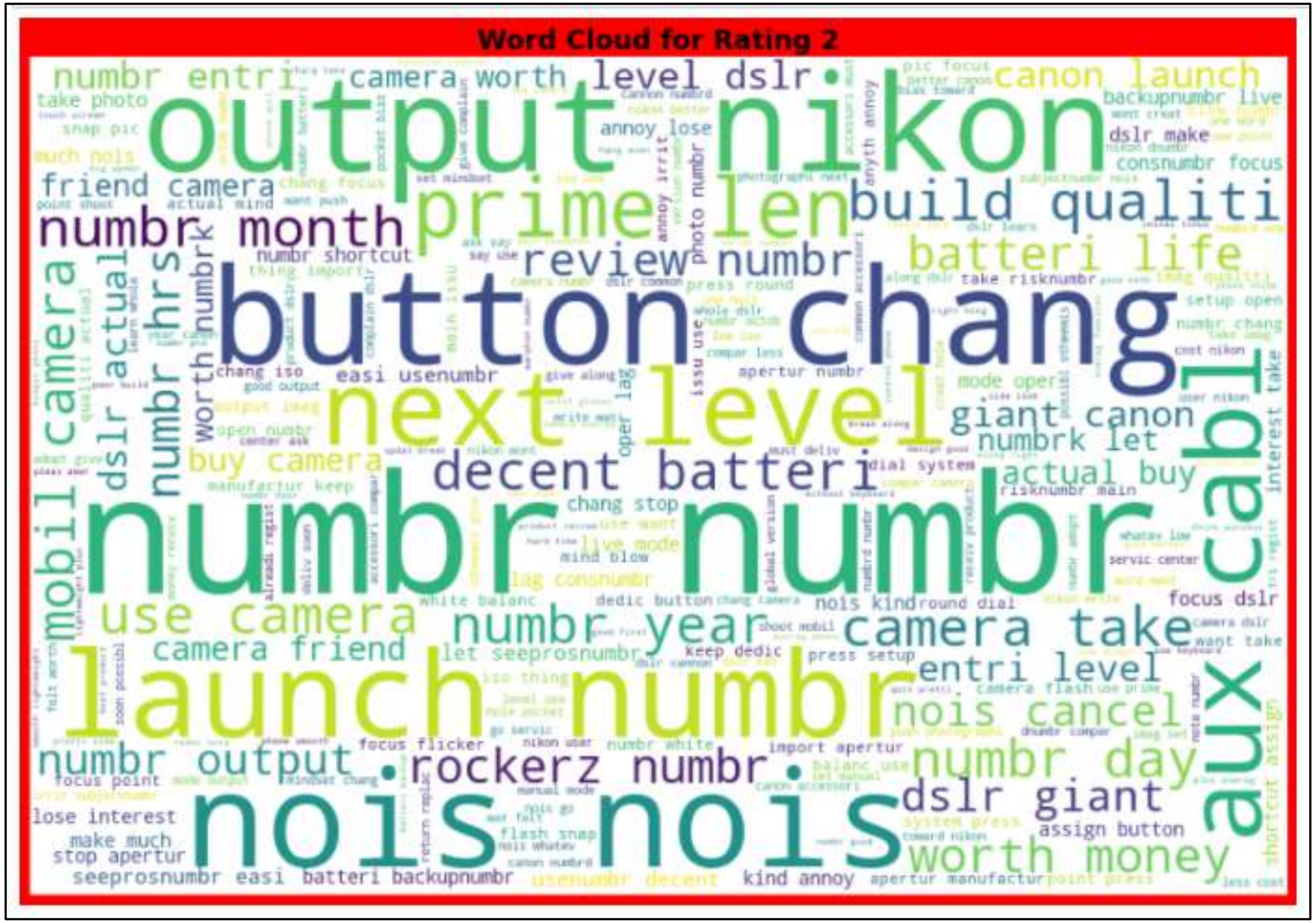
- Convert the text to lowercase
- Remove the punctuations, digits and special characters
- Tokenize the text, filter out the adjectives used in the review and create a new column in data frame
- Remove the stop words
- Stemming and Lemmatising
- Applying Text Vectorization to convert text into numeric

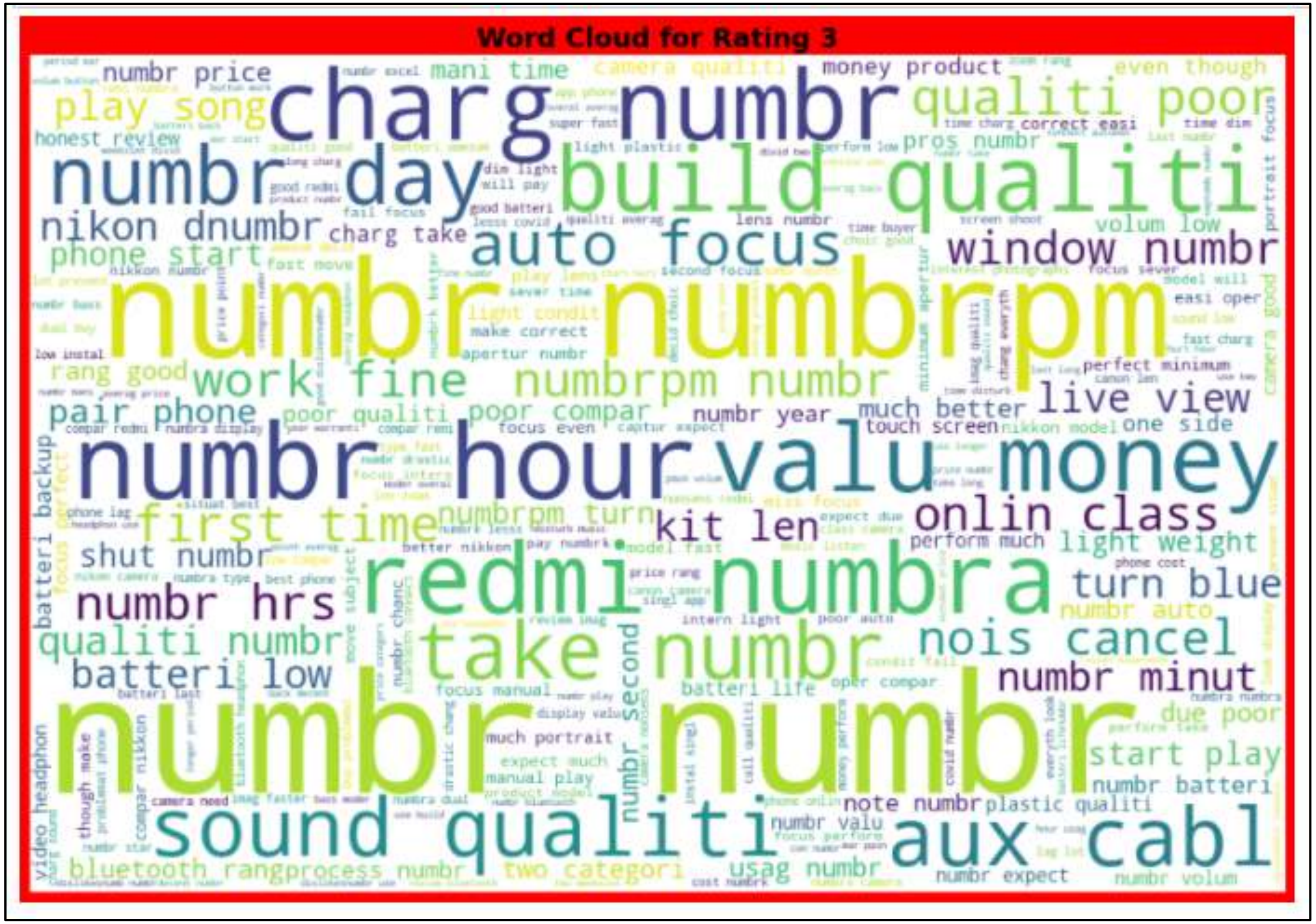
Word Cloud for getting word sense

- Word Cloud is a visualization technique for text data wherein each word is picturized with its importance in the context or its frequency.
- The more commonly the term appears within the text being analysed, the larger the word appears in the image generated.
- The enlarged texts are the greatest number of words used there and small texts are the smaller number of words used.

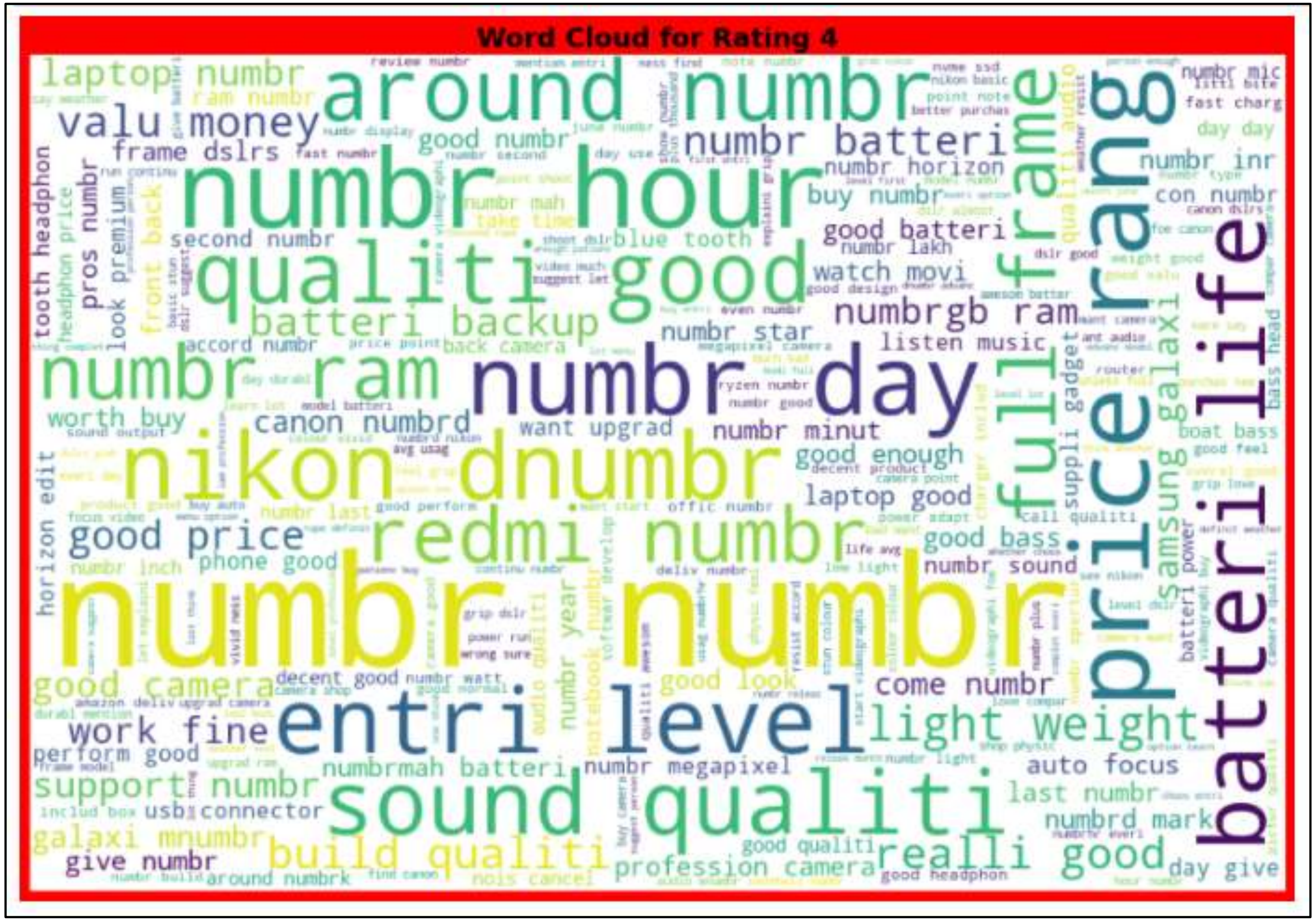


Word Cloud for Rating 1

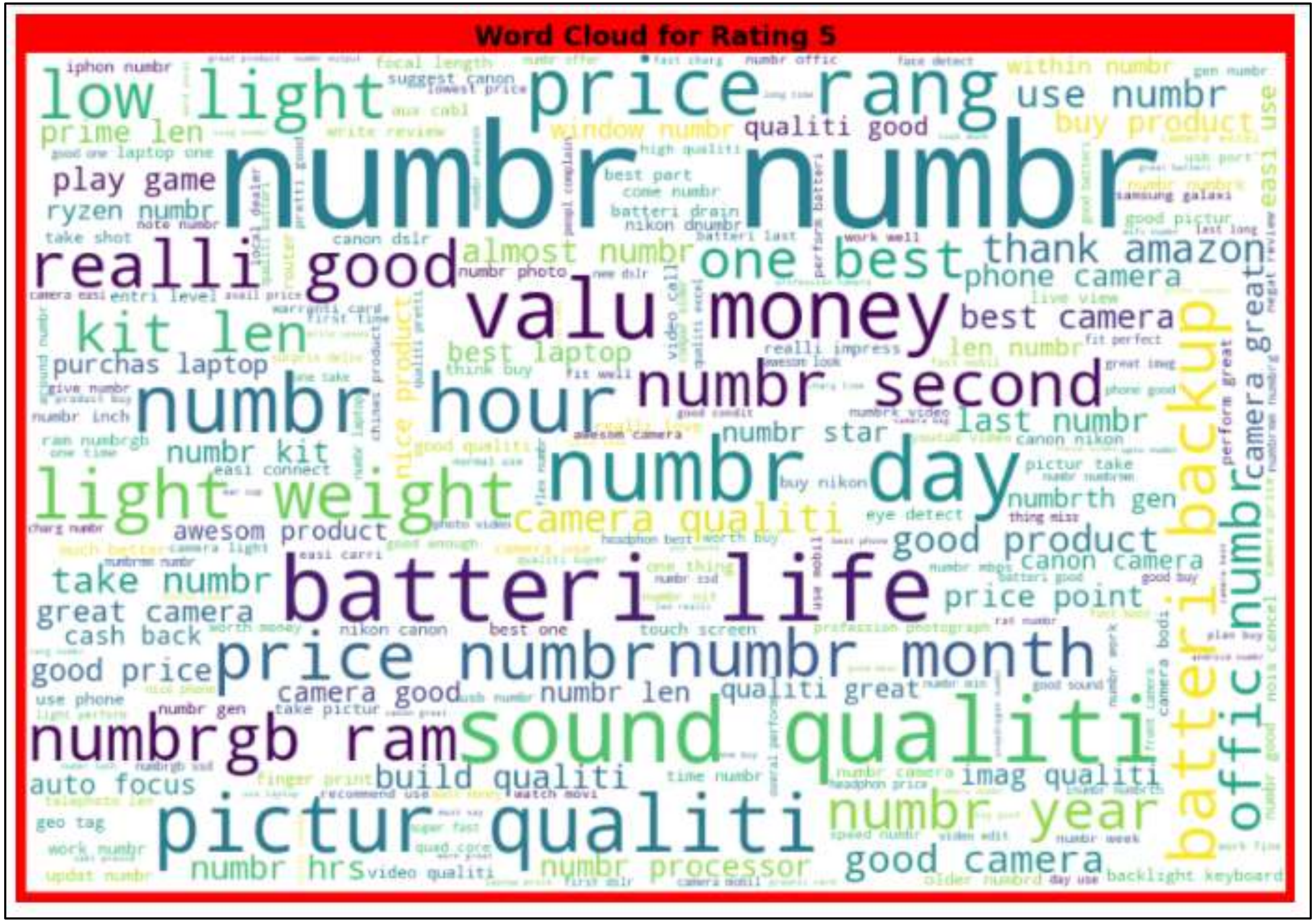

Word Cloud for Rating 2
output nikon
button chang
numbr numbr
launch numbr
nois nois
next level
prime len
aux cabl
build qualiti
numbr month
batteri life
decent batteri
review numbr
mobil camera
use camera
numbr year
camera take
numbr output
rockerz numbr
dslr giant
worth money
numbr day
nois cancel
entri level
giant canon
actual buy
numbrk let
canon launch
level dslr
camera worth
numbr entri
friend camera
camera friend
buy camera
numbr hrs
dslr actual
worth numbrk
take photo
snap pic
much nois
annoy lose
backupnumbr live
dslr make
consnumbr focus
numbr shortcut
easi usenumbr
chang iso
chang stop
live mode
mind blow
lag consnumbr
white balanc
dedic button
nois kind
round dial
focus dslr
let seeprosnumbr
keep dedic
press setup
want take
camera flash
import apertur
focus flicker
nikon user
numbr white
balanc use
assign button
lose interest
make much
stop apertur
focus point
seeprosnumbr easi
batteri backupnumbr
usenumbr decent
kind annoy
apertur manufactur
point press
flash snap
system press
interest take
shortcut assign
setup open
numbr chang
servic center
manufactur keep
open numbr
apertur numbr
press round
take risknumbr
risknumbr main
dial system
mode oper
photo numbr
issu use
oper lag



Word Cloud for Rating 3

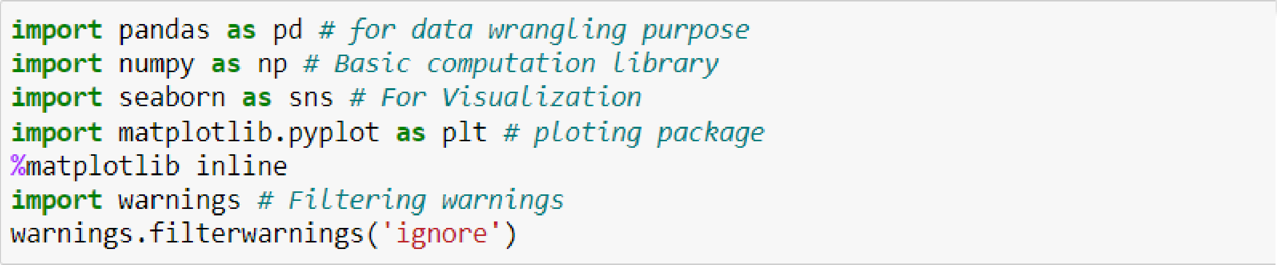


Word Cloud for Rating 4



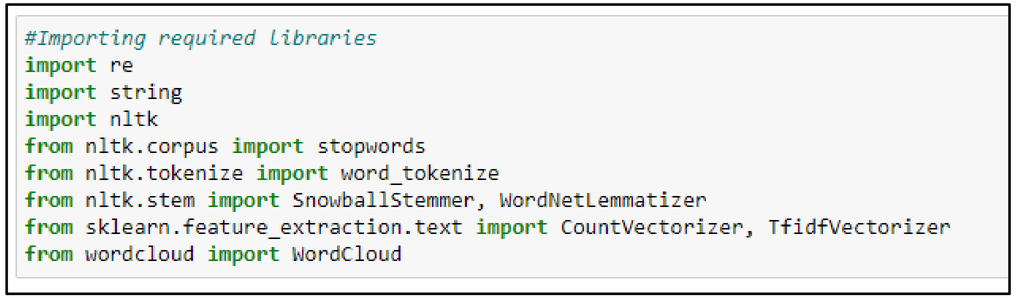
Word Cloud for Rating 5

Web Scraping Library used



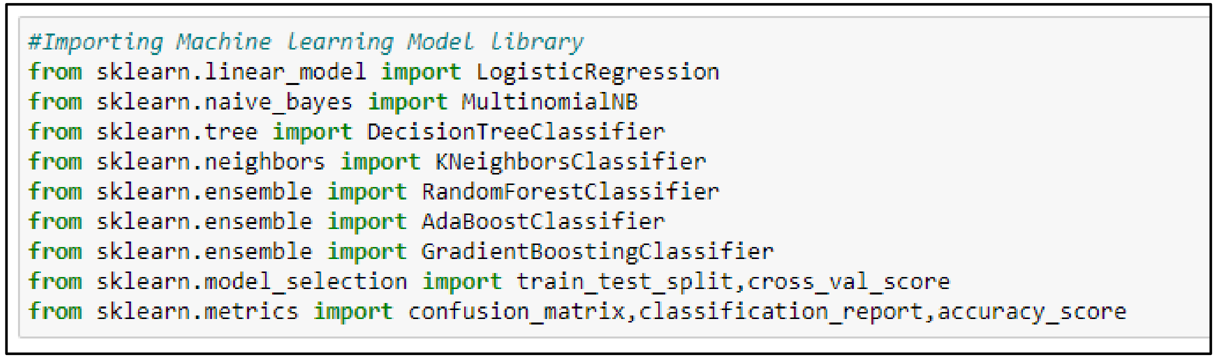
```python
import pandas as pd # for data wrangling purpose
import numpy as np # Basic computation library
import seaborn as sns # For Visualization
import matplotlib.pyplot as plt # ploting package
%matplotlib inline
import warnings # Filtering warnings
warnings.filterwarnings('ignore')
```

Text Mining Library used



```python
#Importing required libraries
import re
import string
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import SnowballStemmer, WordNetLemmatizer
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from wordcloud import WordCloud
```

Machine Learning model building Library used



```python
#Importing Machine Learning Model Library
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import MultinomialNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.model_selection import train_test_split,cross_val_score
from sklearn.metrics import confusion_matrix,classification_report,accuracy_score
```

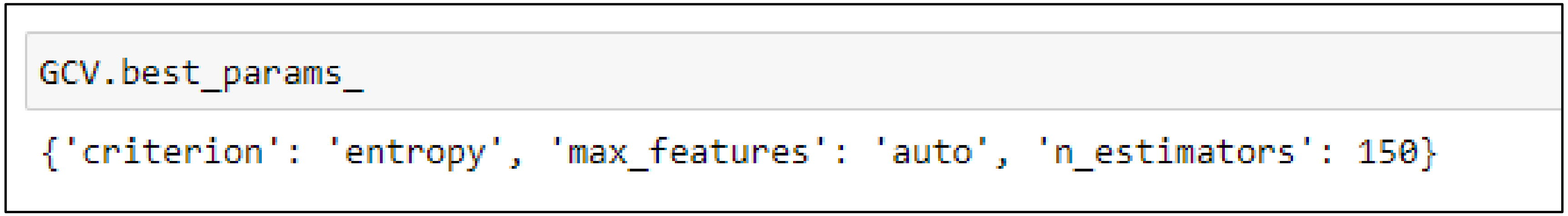
Machine Learning Model Building

The different classification algorithm used in this project to build ML model are as below:

- Random Forest classifier
- Decision Tree classifier
- Logistics Regression
- AdaBoost Classifier
- Gradient Boosting Classifier

Machine Learning Evaluation Matrix

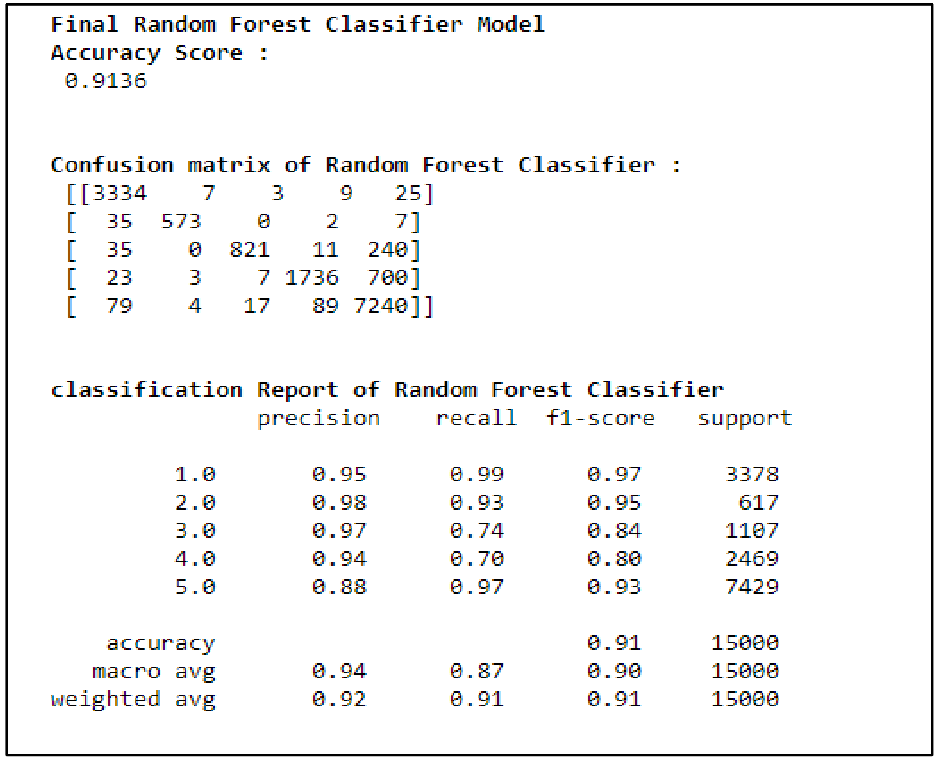
- Random Forest Classifier gives maximum accuracy score.
- Hyper parameter Tuning is perform over this best model.



```
GCV.best_params_
```

```
{'criterion': 'entropy', 'max_features': 'auto', 'n_estimators': 150}
```

Final ML Model



```
Final Random Forest Classifier Model
Accuracy Score :
 0.9136


Confusion matrix of Random Forest Classifier :
 [[3334    7    3    9   25]
 [  35  573    0    2    7]
 [  35    0  821   11  240]
 [  23    3    7 1736  700]
 [  79    4   17   89 7240]]


classification Report of Random Forest Classifier
               precision    recall  f1-score   support

         1.0       0.95      0.99      0.97      3378
         2.0       0.98      0.93      0.95       617
         3.0       0.97      0.74      0.84      1107
         4.0       0.94      0.70      0.80      2469
         5.0       0.88      0.97      0.93      7429

    accuracy                           0.91     15000
   macro avg       0.94      0.87      0.90     15000
weighted avg       0.92      0.91      0.91     15000
```

5-fold Cross validation performed over all model. We can see that Random Forest Classifier gives us good Accuracy and maximum f1 score along with best Cross-validation score. Hyperparameter tuning is applied over Random Forest model and used it as final model.

Machine Learning Evaluation Matrix

Algorithm	Accuracy Score	Recall	Precision	F1 Score	CV Score
Logistics Regression	0.9071	0.86	0.94	0.91	0.5794
Decision Tree Classifier	0.8957	0.86	0.90	0.90	0.5298
Random Forest Classifier (RFC)	0.9133	0.87	0.94	0.91	0.5621
Gradient Boosting Classifier	0.9022	0.86	0.94	0.90	0.6113
Ada Boost Classifier	0.5932	0.39	0.60	0.59	0.5204
Final Model (RFC- Tuned)	0.9136	0.87	0.94	0.91	0.5730

CONCLUSION

- Successfully developed Machine learning model to predict product review ratings.
- NLTK library used for text Mining.
- Random forest classifier model is best model with accuracy score of 91.36%.