RATING PREDICTION PROJECT

Submitted by:

ANISH ANTONY

ABSTRACT:

- For every company, it needs to understand the customer’s feedback. So based on the review it can update its products. The main objective is text process the review and perform language text processing thereby predicting the rating with the review text. This can be done using various language processing techniques such as Count Vectorizer and TFIDF with machine learning algorithms

- Keywords: Review, Rating, Selenium, Data cleaning, Count Vectorizer, Naïve Bayers

The features used in the dataset are:

- Rating
- Review

Data Cleaning:

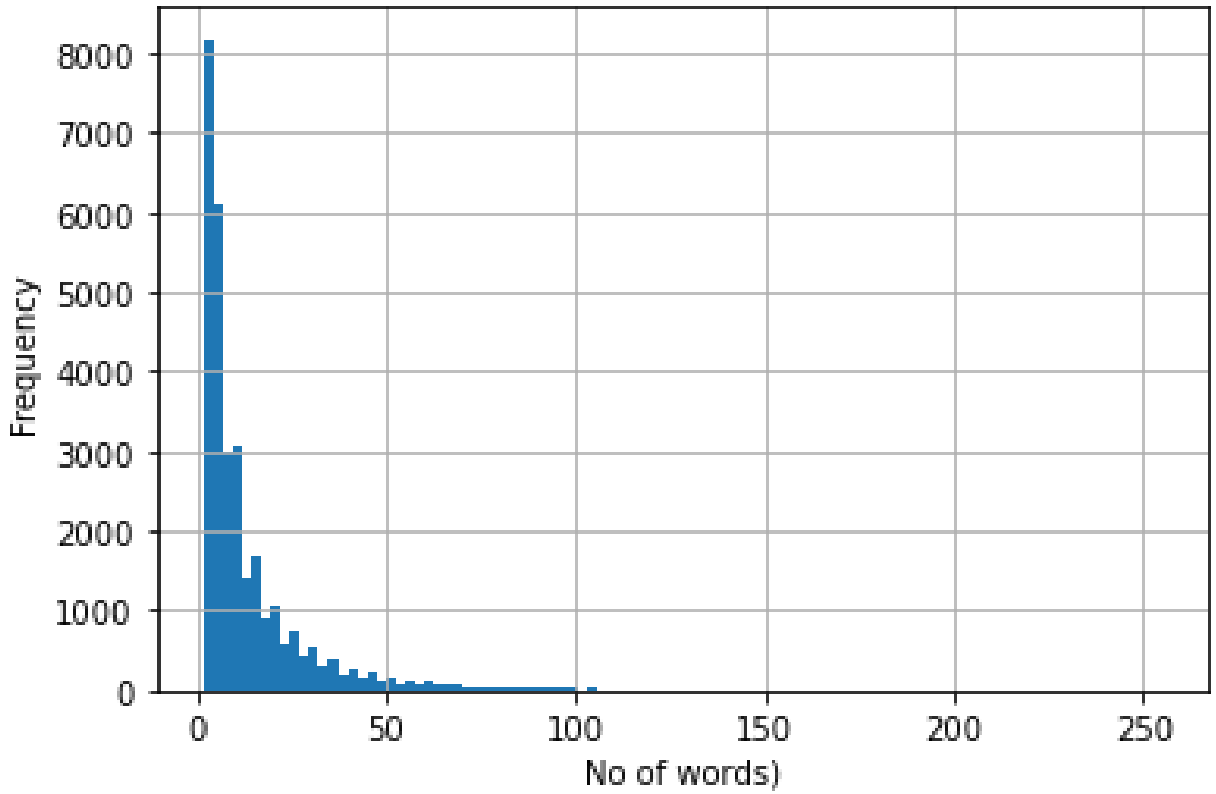
Categorize the excel as flipkart and amazon, since they have different formats

1. Flipkart:
   - Extract the first string and store it as Rating
   - Extract the remaining strings and store it as review thereby removing the name ,location , etc/
2. Amazon:
   - Ratings are extracted separately during web scrapping
   - Extract the remaining strings and store it as review thereby removing the name ,location , etc/

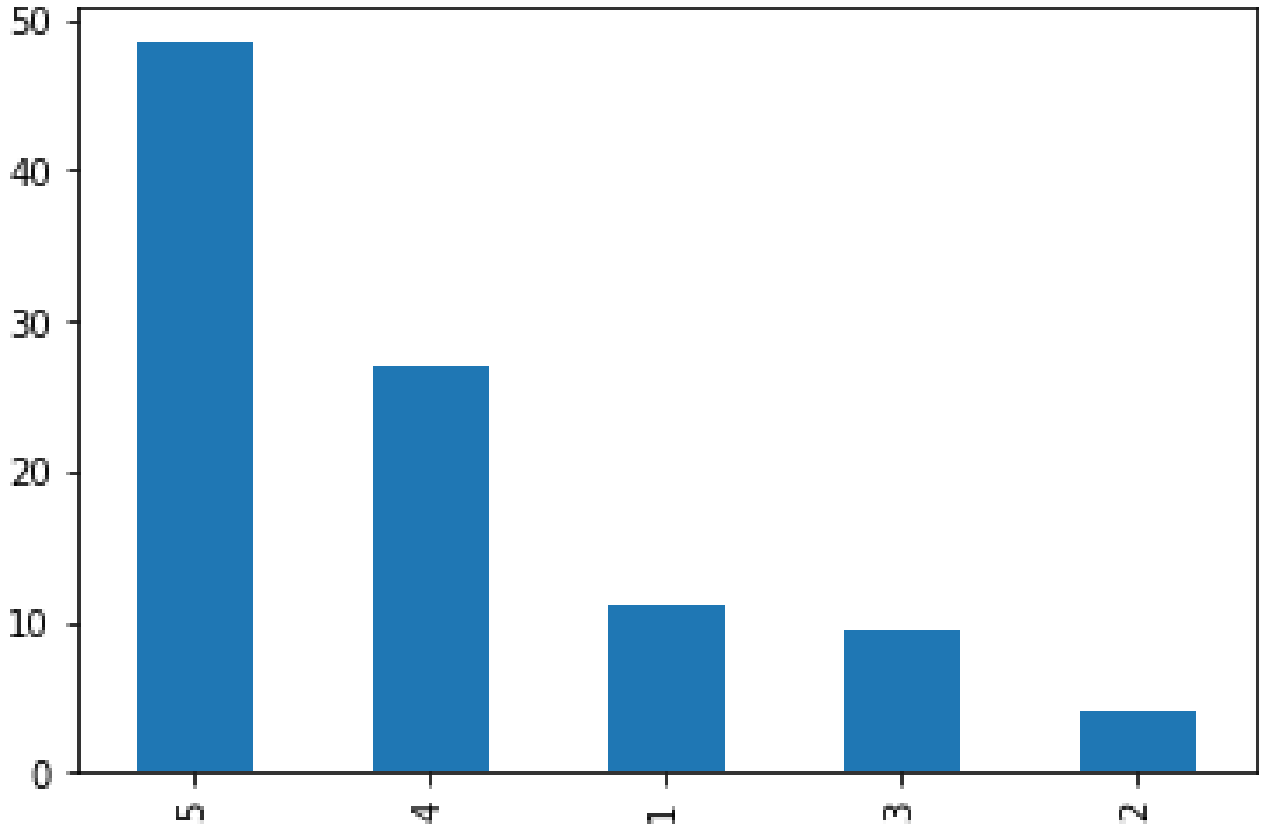
Hardware and Software Requirements and Tools Used:

- Hardware – PC Windows 10, 4 GB Ram
- Software – Google chrome, MS Excel, Python, Selenium webdriver
- Libraries – Pandas, NumPy, Matplotlib, Seaborn, sklearn, SciPy. Stats
    - ❑ Browsing – Google Chrome
    - ❑ Webscraping – Python, Selenium webdriver
    - ❑ Data cleaning – Python, Pandas, NumPy & SciPy. Stats
    - ❑ Data visualization – Matplotlib & Seaborn
    - ❑ Machine learning – Sklearn

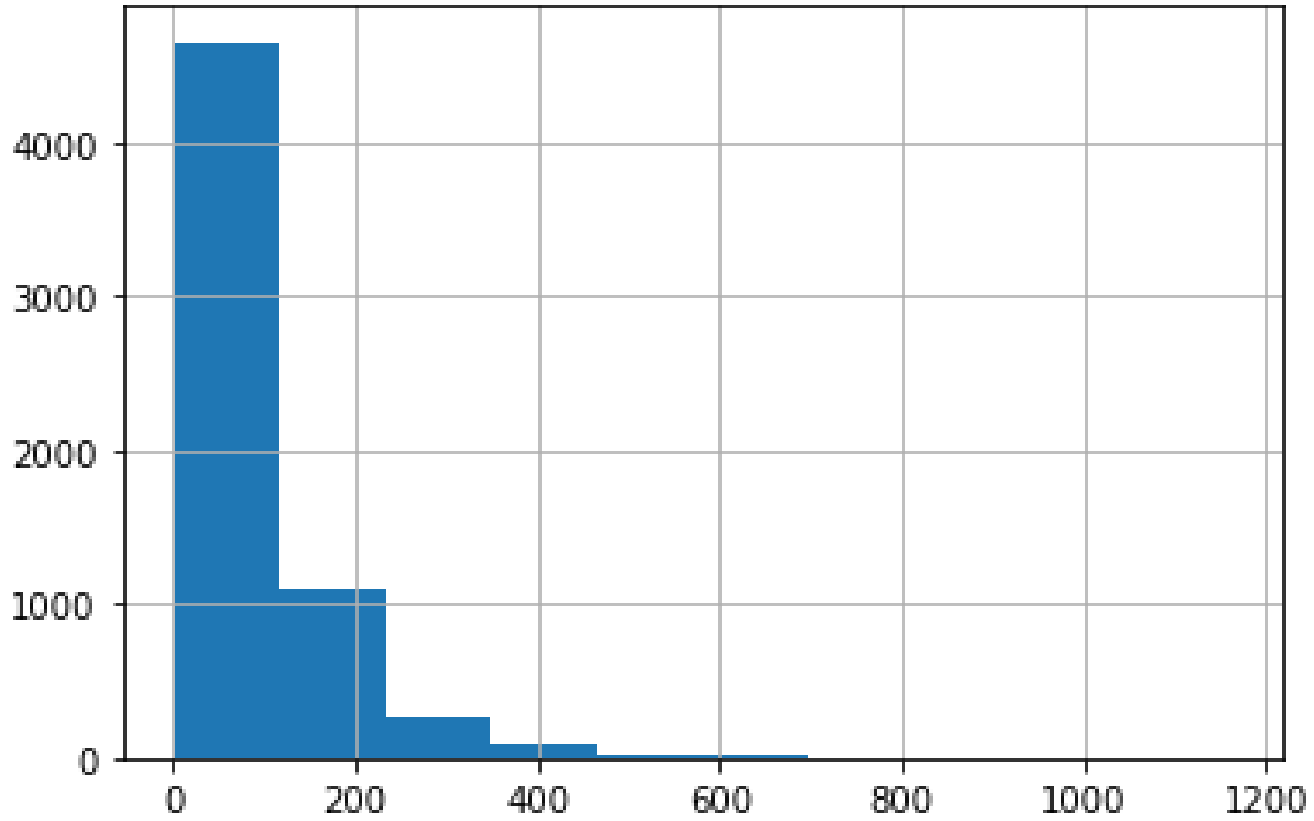
No of words vs Frequency:



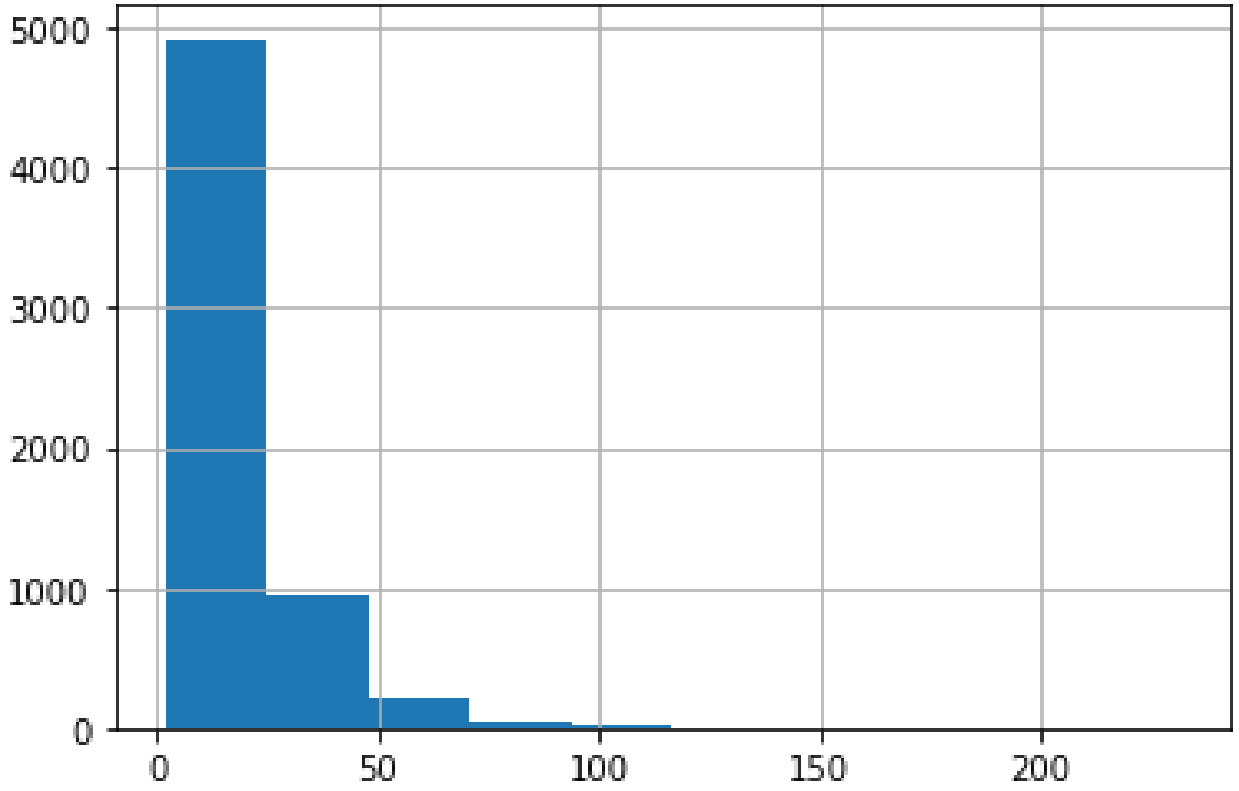
Percentage of Ratings:



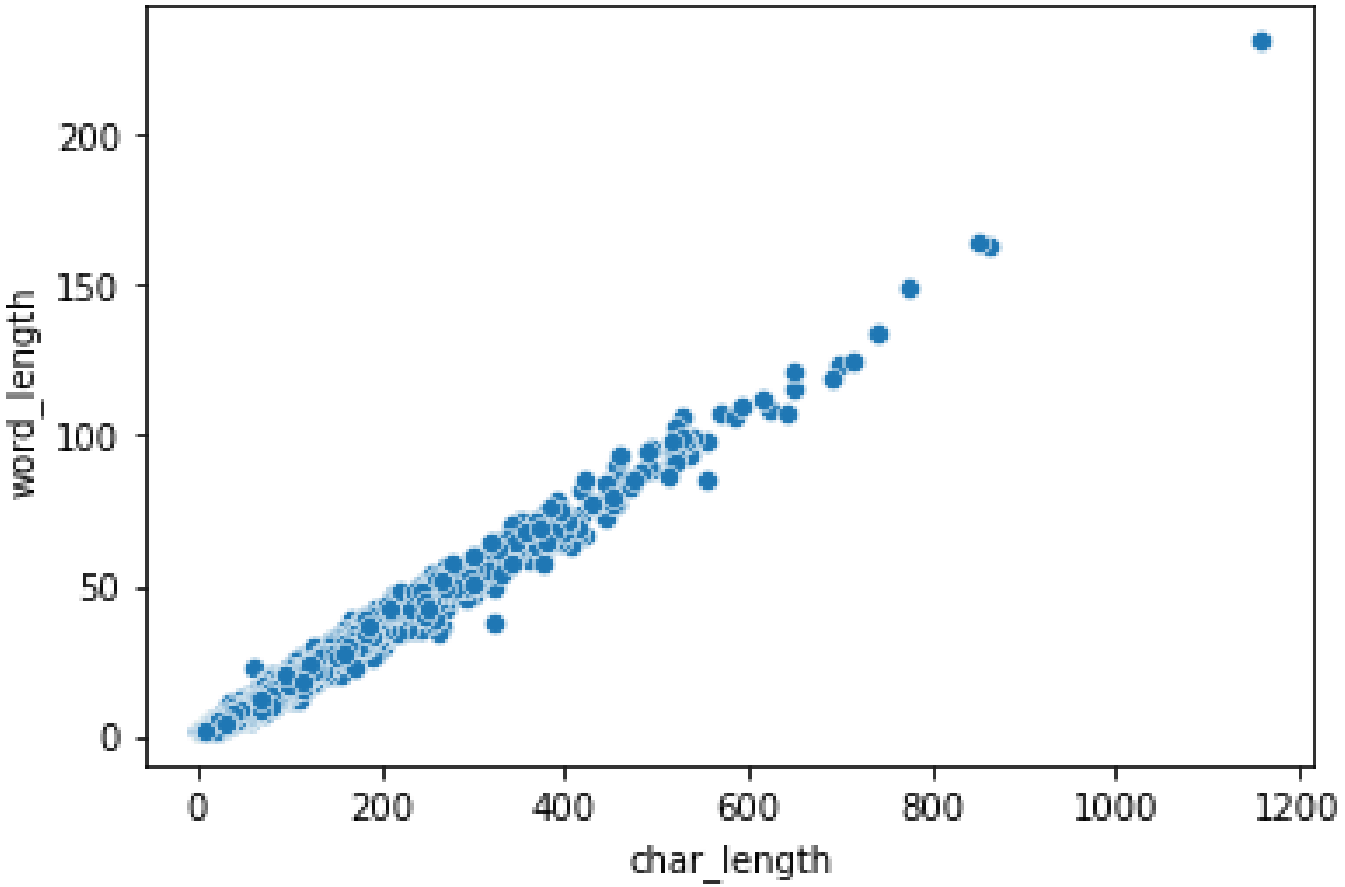
Character Length:



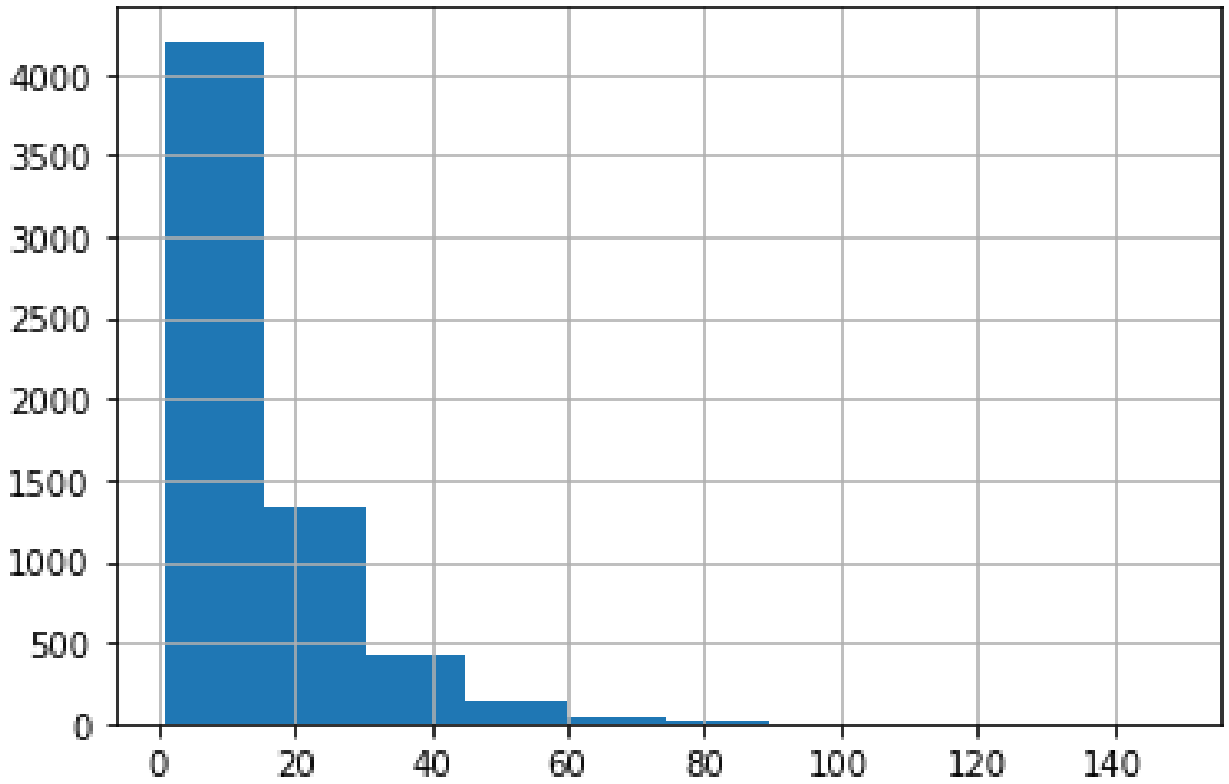
Word Length:



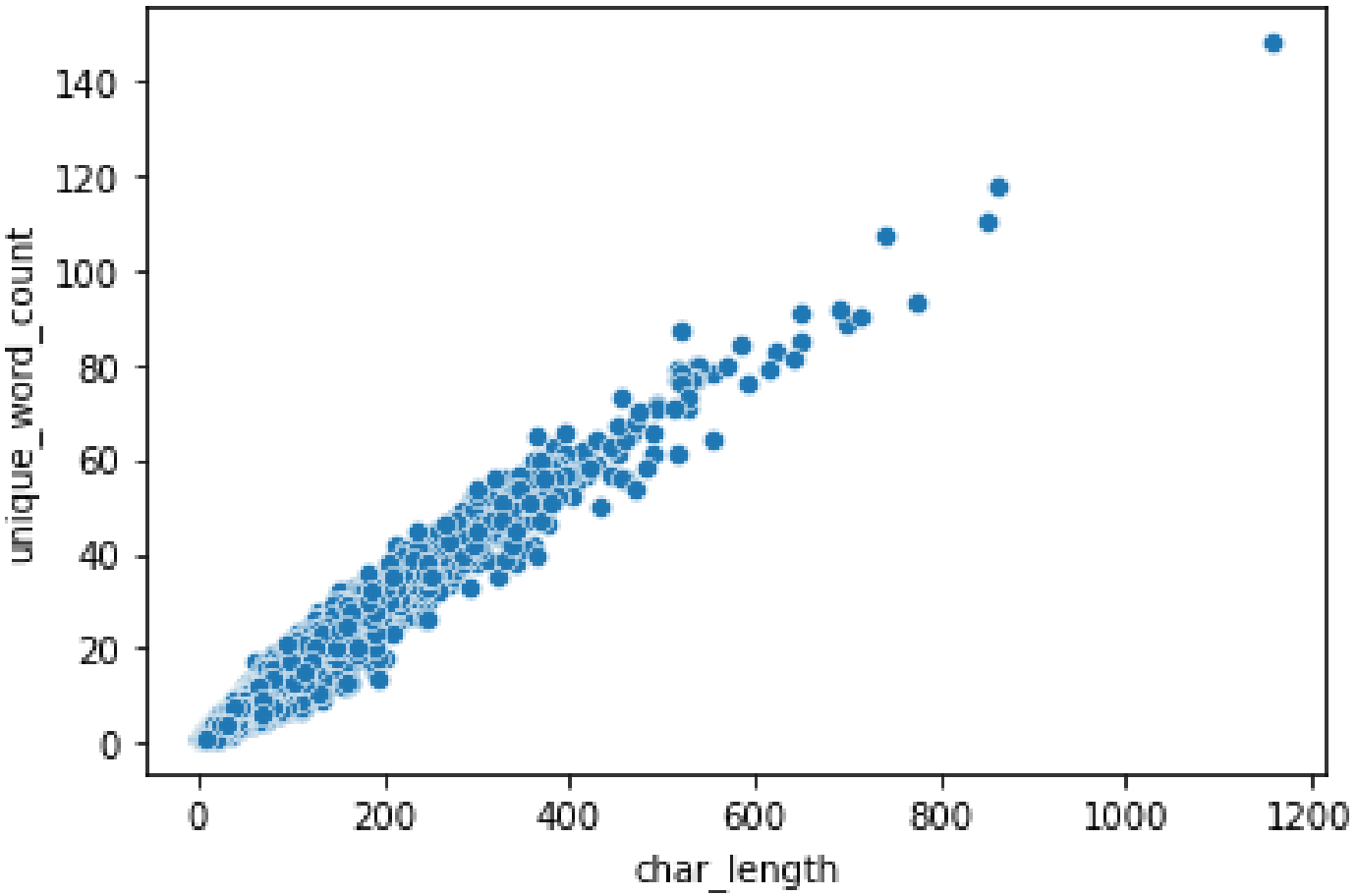
Word Length vs Character Length:



Unique Word:


4000
3500
3000
2500
2000
1500
1000
500
0
0
20
40
60
80
100
120
140

Unique Word vs Character Length:



Word cloud:


hit
don t
return
card
sound quality
use
sound
Good product
time
using
bass
s
apple
still
sturdy
ok
fun
better
well
even
one
work
people found
charge
low
charger
service
original
price
good
quality
working
warranty
Amazing
play
found
Overall
Amazon
light
need
buy
go
best
good quality
issue
will
fast
awesome
Excellent
item
mosquitoe
hour
game
product
earphone
person
iPhone Nice product
battery life
bought
brand
week
Please
charging
One person
worth
really
poor
bad
got
phone
expected
now
UNO
bat
last
month
ear
call
problem
money
day
used
kid
waste
year
family
purchase
Value
want
great
stopped working
Nice
battery
received
racquet
much
durable
look
replacement
Worst

Testing of Identified Approaches (Algorithms):

- Random Forest Classifier
- Linear SVC
- Logistic Regression
- Multinomial NB
- Bernoulli NB
- LGBM Classifier
- SGD Classifier

Count Vectorizer:

- Accuracy Score: 49.8371335504886

CLASSIFICATION REPORT :

	precision	recall	f1-score	support
1	0.48	0.60	0.53	242
2	0.58	0.43	0.49	230
3	0.42	0.31	0.36	258
4	0.40	0.40	0.40	258
5	0.60	0.78	0.68	240
accuracy			0.50	1228
macro avg	0.50	0.50	0.49	1228
weighted avg	0.49	0.50	0.49	1228

Confusion Matrix:

```
Confusion Matrix :
 [[144  20  37  30  11]
 [ 55  98  36  29  12]
 [ 67  28  80  57  26]
 [ 31  15  32 103  77]
 [  2   8   7  36 187]]
```

Tfidf Vectorizer:

- Accuracy Score: 48.9413680781759

CLASSIFICATION REPORT :

	precision	recall	f1-score	support
1	0.48	0.61	0.54	242
2	0.54	0.46	0.49	230
3	0.40	0.29	0.34	258
4	0.40	0.36	0.38	258
5	0.60	0.75	0.66	240
accuracy			0.49	1228
macro avg	0.48	0.49	0.48	1228
weighted avg	0.48	0.49	0.48	1228

Confusion Matrix:

```
Confusion Matrix :
 [[148  22  37  25  10]
 [ 54 105  37  23  11]
 [ 69  40  75  50  24]
 [ 34  19  35  93  77]
 [  2  10   5  43 180]]
```

Key Findings and Conclusions of the Study:

- This dataset has been taken from 2 websites of this, Flipkart and Amazon constitutes the majority of data
- Since the target feature is categorical data, this problem can be solved by classification algorithms
- Count Vectorizer gives an accuracy score . 49.8371335504886
- TFIDF gives an accuracy score . 48.9413680781759