Project – Sentimental Analysis on Amazon Data (P116)

Group 6 - Tripti, Rushikesh, Adarsh, Mukund, Harshitha, Ajinkya

Guided by - Parth Sagar

Project start date - 05/05/2022

Business Problem

To perform Sentimental Analysis on a Amazon product

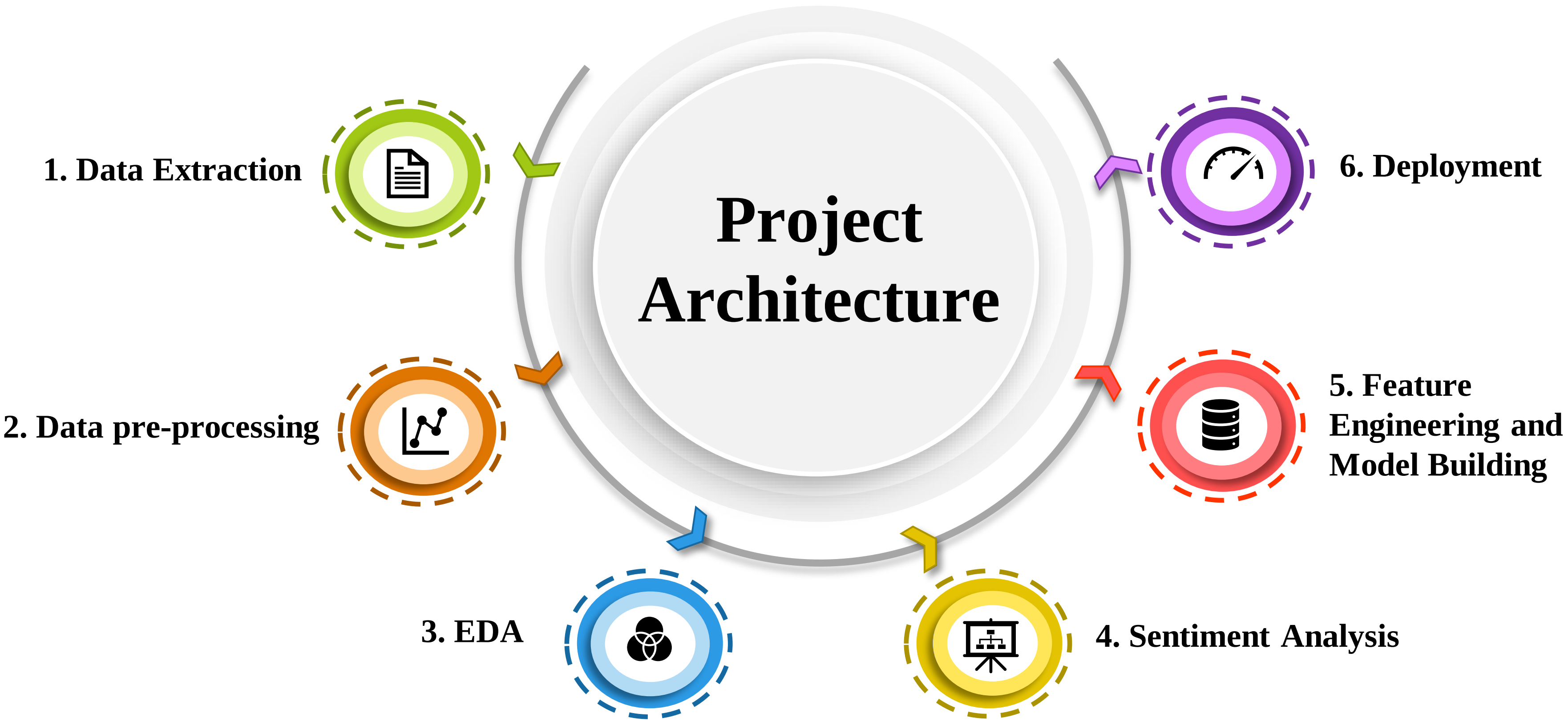
Objective

The objective of the analysis is to get daily Analysis of a product such as emotions, sentiment etc. using Amazon data of our choice.

Data Insights

The product chosen for our Data extraction for Sentimental Analysis is Redmi Note 8. The dataset contains over 3000 product reviews from Amazon.com along with their corresponding rating stars and usernames of the customers .

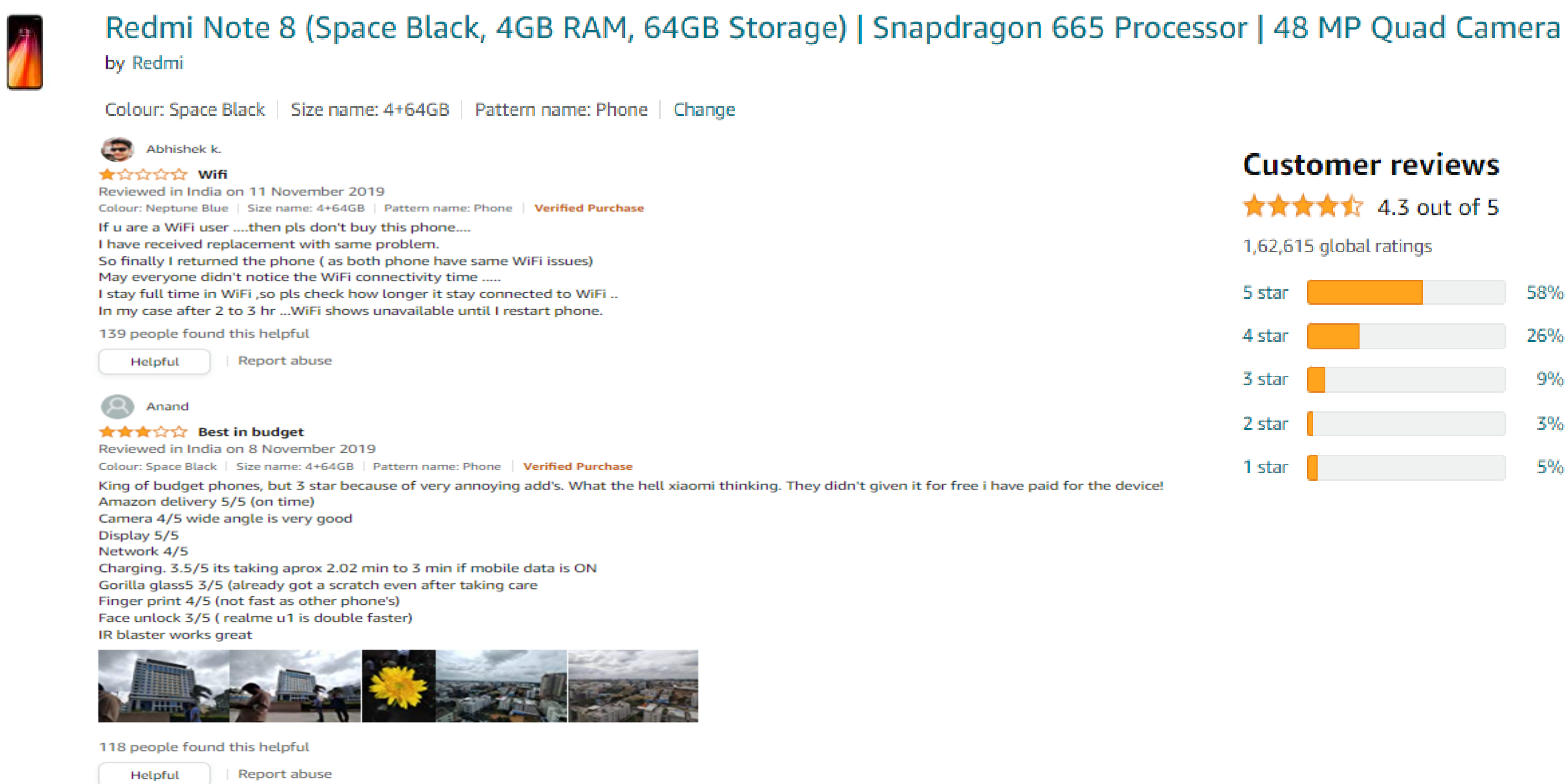
Project Flow



Amazon Product chosen for Sentiment Analysis

Product name – Redmi Note 8 (Space Black, 4GB RAM, 64GB Storage)

Product URL - https://www.amazon.in/Redmi-Note-Space-Black-Storage/product-reviews/B07X4PXKZ7/ref=cm_cr_dp_d_show_all_btm?ie=UTF8&reviewerType=all_reviews



Data Extraction

The data was extracted using Beautiful Soup Tool

Number of ratings – 1,62,615
Number of reviews – 42,084

Number of extracted reviews and ratings - 3100

	stars	reviews	name
0	1	The media could not be loaded.\n ...	Ashraf
1	5	Febulas performance Redmi Note 8 ...I love it ...	Anil kumar sharma
2	5	best mobile under 10000	Mahendra
3	5	Redmi note 8 is the best Smartphone under 10k ...	Shah Arsalan
4	5	Loving the phone....Purchased with bank discou...	R.T
5	5	Pros- batteryCameraPriceLookCons- delicate. So...	Sadhna agrawal
6	5	Excellent phone under 10,000. Specialy 18 watt...	Amazon Customer
7	2	Redmi has been a prominent smartphone brand wh...	Aman Singh
8	5	Excellent phone in this price point	Rao
9	5	I purchased it from redmi store. Its performan...	Arjun
10	5	Nice product ...India's No One. Brand	VaLLaRaSu V
11	1	If u are a WiFi user ....then pls don't buy th...	Abhishek k.
12	3	King of budget phones, but 3 star because of v...	Anand
13	5	This product is very good and very good qualit...	vikash kumar bharti
14	1	Camera not good. Bad performance . Battery ch...	Nikita
15	5	You may miss such like a worthy mobile in just...	C.Suresh
16	5	I got this device on 1st November 1 week befor...	Kushal Dutta
17	5	It took around 9 days to get the product deliv...	Naeem Abbas
18	5	Purple is too good.	Amazon Customer
19	1	Very Bad experience with this phone quality ...	siddharth

Data Pre-processing

Steps followed for pre-processing -

Data Cleaning

- Converting text to lowercase
- Removing punctuations
- Removing stopwords
- Removing accents
- Removing empty text spaces
- Removing hyperlinks

Step 1

Tokenization

A way of separating a piece of text into smaller units called tokens

Step 2

POS(part of speech) Labelling

The tag in case of is a part-of-speech tag, and signifies whether the word is a noun, adjective, verb

Step 3

Lemmatization

Normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma

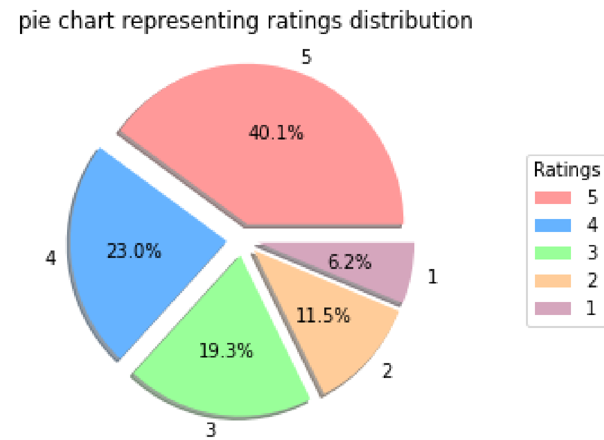
Step 4

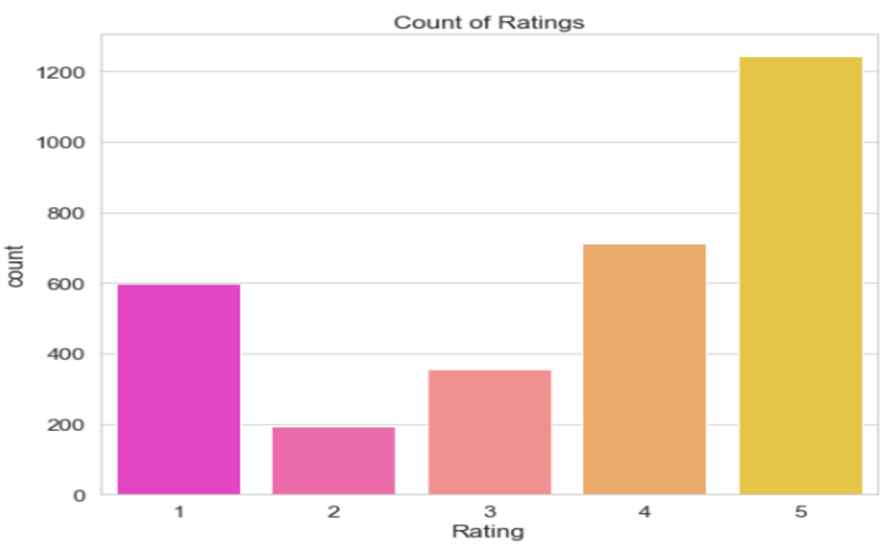
	stars	reviews	cleaned_reviews	tokens	POS_tagging	Lemmas
0	1	The media could not be loaded.\n ...	media could loaded phone hanged many times ret...	[media, could, loaded, phone, hanged, many, ti...	[(media, n), (could, None), (loaded, v), (phon...	medium could load phone hang many time retur...
1	5	Febulas performance Redmi Note 8 ...I love it ...	febulas performance redmi note 8 love first ti...	[febulas, performance, redmi, note, 8, love, f...	[(febulas, n), (performance, n), (redmi, v), (...	febulas performance redmi note 8 love first ...
2	5	best mobile under 10000	best mobile 10000	[best, mobile, 10000]	[(best, r), (mobile, a), (10000, None)]	best mobile 10000
3	5	Redmi note 8 is the best Smartphone under 10k ...	redmi note 8 best smartphone 10k year 2019	[redmi, note, 8, best, smartphone, 10k, year, ...	[(redmi, a), (note, n), (8, None), (best, a), ...	redmi note 8 best smartphone 10k year 2019
4	5	Loving the phone....Purchased with bank discou...	loving phone purchased bank discount 6gb 128gb...	[loving, phone, purchased, bank, discount, 6gb...	[(loving, v), (phone, n), (purchased, v), (ban...	love phone purchase bank discount 6gb 128gb ...
5	5	Pros- batteryCameraPriceLookCons- delicate. So...	pros batterycamerapricelookcons delicate handl...	[pros, batterycamerapricelookcons, delicate, h...	[(pros, n), (batterycamerapricelookcons, n), (...	pro batterycamerapricelookcons delicate hand...
6	5	Excellent phone under 10,000. Specialy 18 watt...	excellent phone 10 000 specialy 18 watt fast c...	[excellent, phone, 10, 000, specialy, 18, watt...	[(excellent, a), (phone, n), (10, None), (000,...	excellent phone 10 000 specialy 18 watt fast...
7	2	Redmi has been a prominent smartphone brand wh...	redmi prominent smartphone brand given quality...	[redmi, prominent, smartphone, brand, given, q...	[(redmi, n), (prominent, a), (smartphone, n), ...	redmi prominent smartphone brand give qualit...
8	5	Excellent phone in this price point 👌👌👌	excellent phone price point	[excellent, phone, price, point]	[(excellent, a), (phone, n), (price, n), (poin...	excellent phone price point
9	5	I purchased it from redmi store. Its performan...	purchased redmi store performance awesome supe...	[purchased, redmi, store, performance, awesome...	[(purchased, v), (redmi, n), (store, n), (perf...	purchase redmi store performance awesome sup...

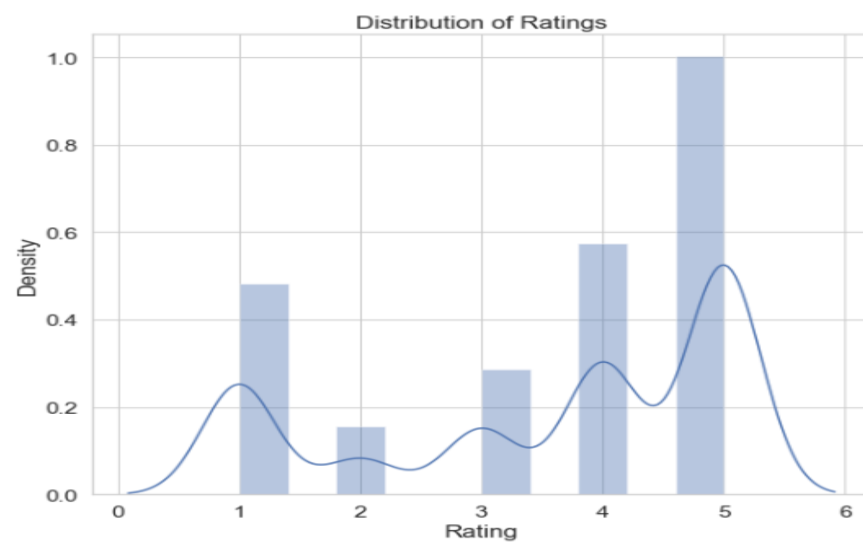
EXCELR
Raising Excellence

Exploratory Data Analysis (EDA)

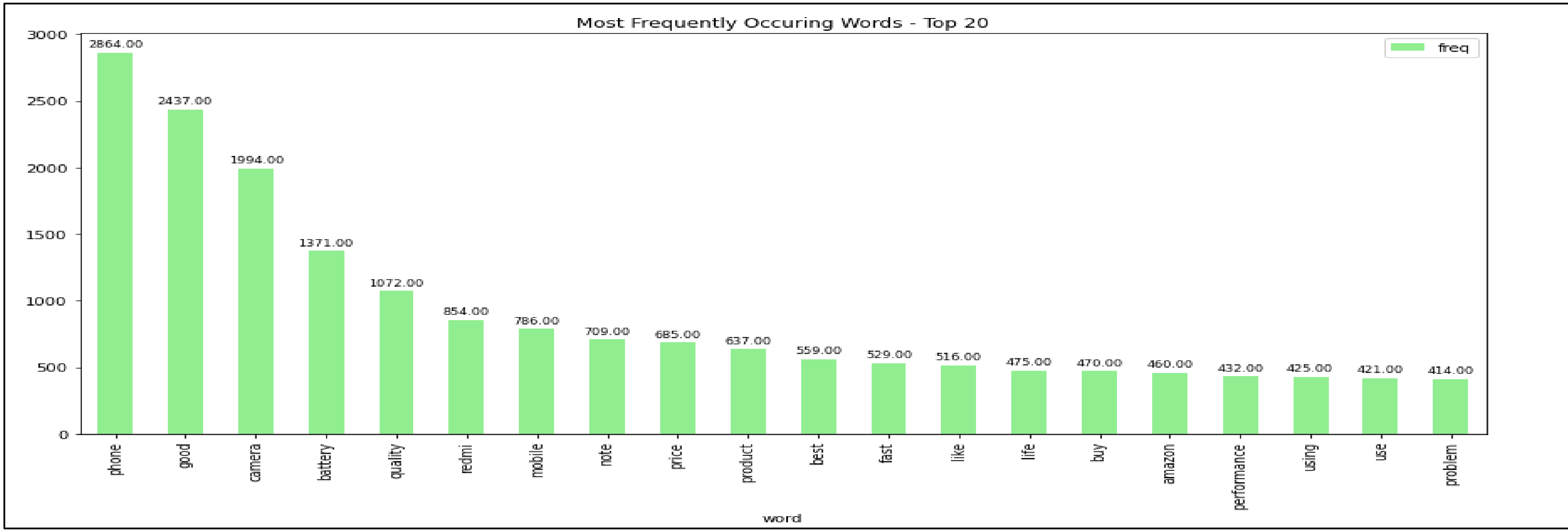
- 40.1% ratings are positive reviews with 5 stars
- Whereas the reviews with 4 stars or less contribute to 60% of total ratings.
- On an average the reviews are having ratings around 3.5 to 4.5 marking up to combined total of 42.3% of the total ratings and only 17.7% of the total reviews are rated between 1 and 2 star.

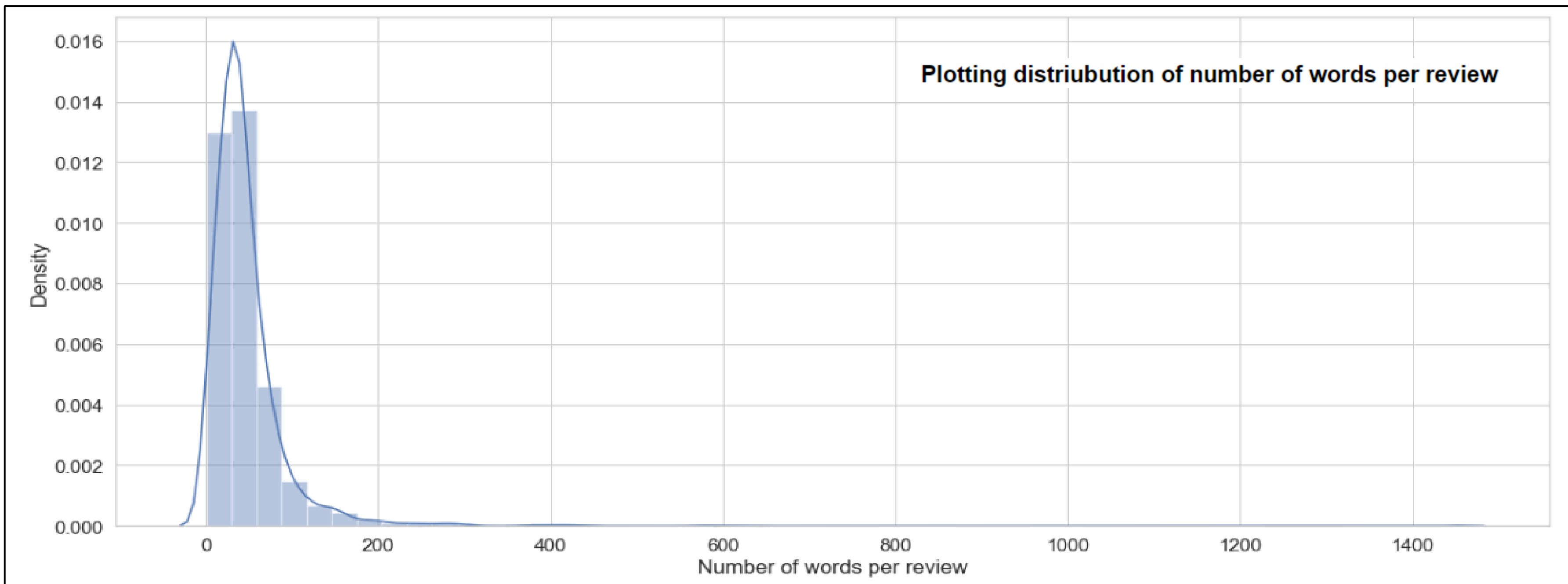


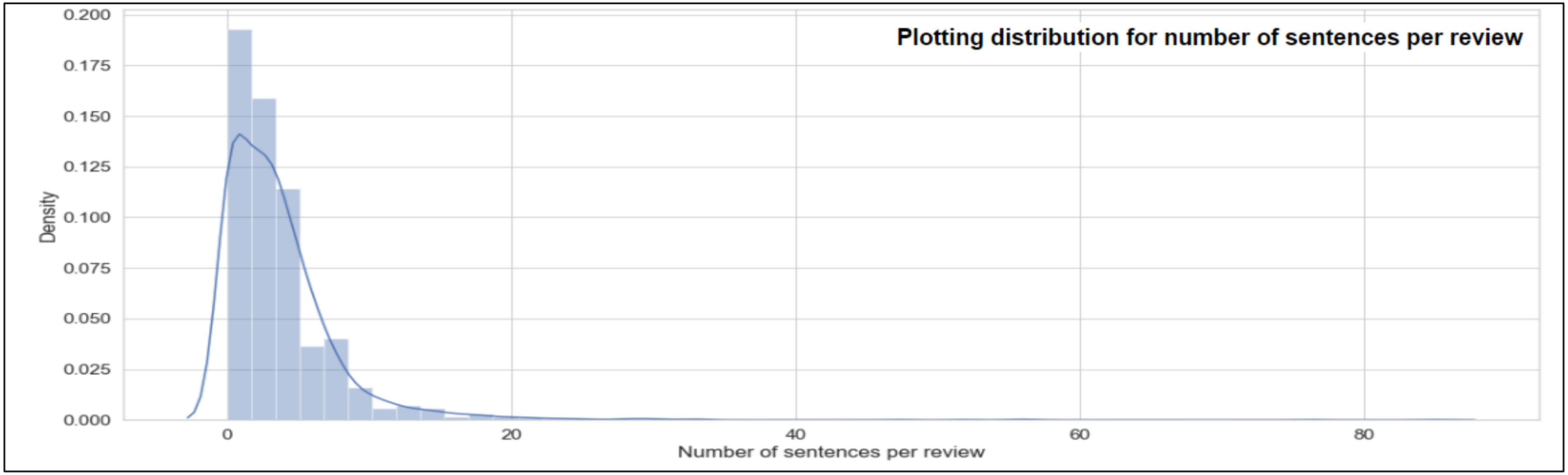


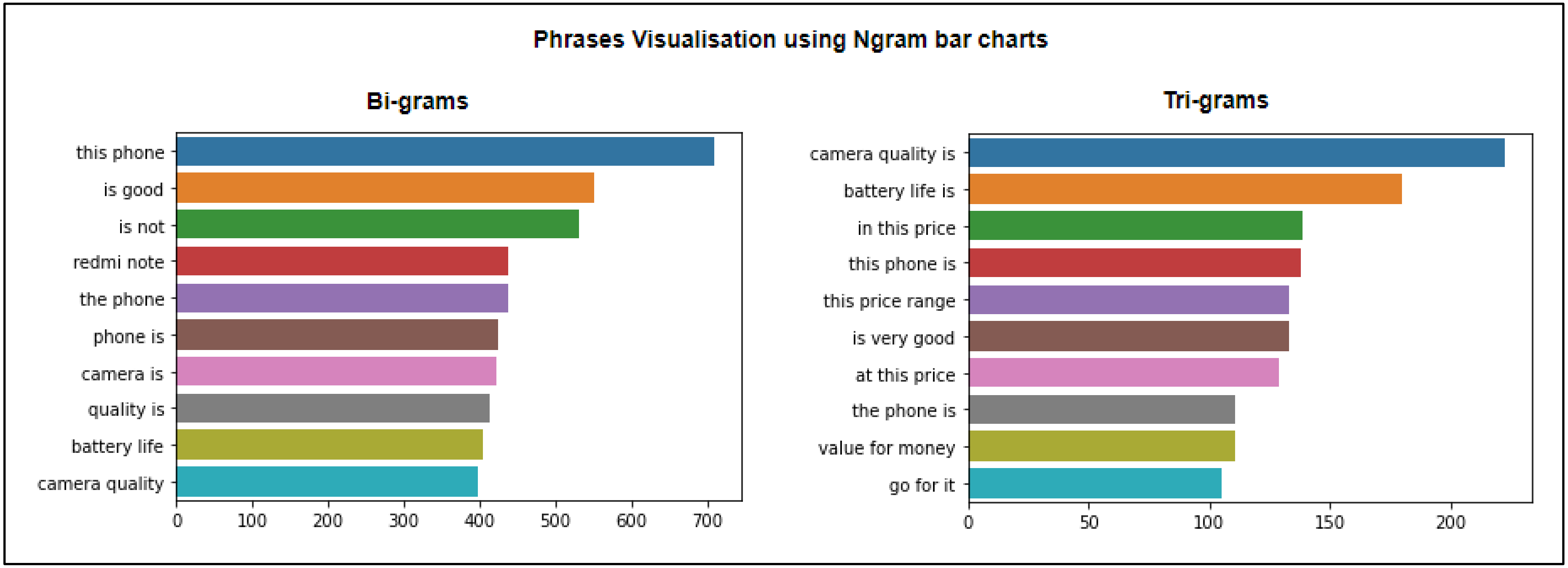


Plotting top 20 most frequent occuring words






Plotting distribution for number of sentences per review
Density
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
20
40
60
80
Number of sentences per review


Phrases Visualisation using Ngram bar charts
Bi-grams
this phone
is good
is not
redmi note
the phone
phone is
camera is
quality is
battery life
camera quality
0
100
200
300
400
500
600
700
Tri-grams
camera quality is
battery life is
in this price
this phone is
this price range
is very good
at this price
the phone is
value for money
go for it
0
50
100
150
200

Sentimental Analysis

Using VADER SentimentIntensityAnalyser to calculate the sentiment score

VADER (Valence Aware Dictionary for Sentiment Reasoning) is a model used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion

	stars	reviews	cleaned_reviews	tokens	POS_Tagging	Lemmas	sentiment_score
3095	5	Very good and premium look. Excellent sound qu...	good premium look excellent sound quality fast...	[good, premium, look, excellent, sound, qualit...	[(good, a), (premium, n), (look, n), (excellen...	good premium look excellent sound quality fa...	0.93
3096	4	It is gud phone while playing heavy graphics g...	gud phone playing heavy graphics game phone he...	[gud, phone, playing, heavy, graphics, game, p...	[(gud, n), (phone, n), (playing, v), (heavy, a...	gud phone play heavy graphic game phone heat...	0.57
3097	5	Gohead buy it guys.	gohead buy guys	[gohead, buy, guys]	[(gohead, a), (buy, n), (guys, n)]	gohead buy guy	0.00
3098	5	Good product	good product	[good, product]	[(good, a), (product, n)]	good product	0.44
3099	4	Good	good	[good]	[(good, a)]	good	0.44

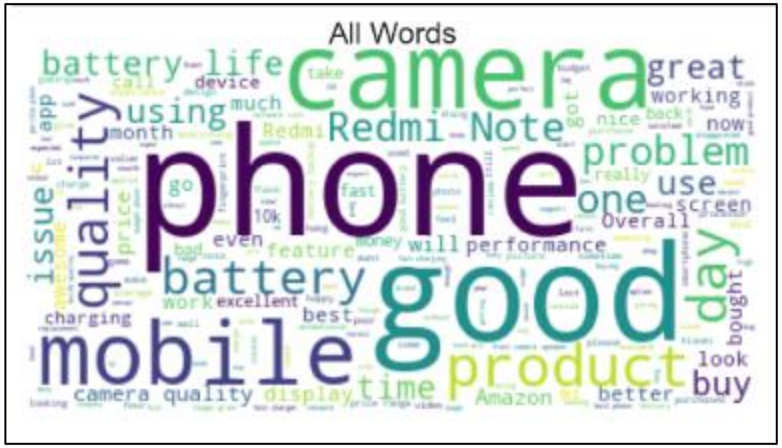
- Polarity Classification(scores) – since we are classifying the reviews as either positive and negative the polarity scores will help us know the positivity rate and negativity rate of the reviews

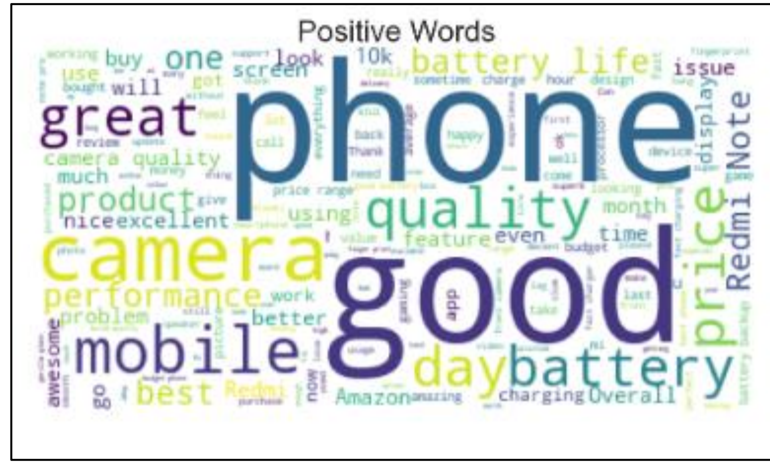
- The Compound score(comp_score) is a metric that calculates the sum of all the lexicon ratings which have been normalized between -1(most extreme negative) and +1 (most extreme positive).
  positive sentiment : (compound score >= 0.05)
  neutral sentiment : (compound score > -0.05) and (compound score < 0.05)
  negative sentiment : (compound score <= -0.05)

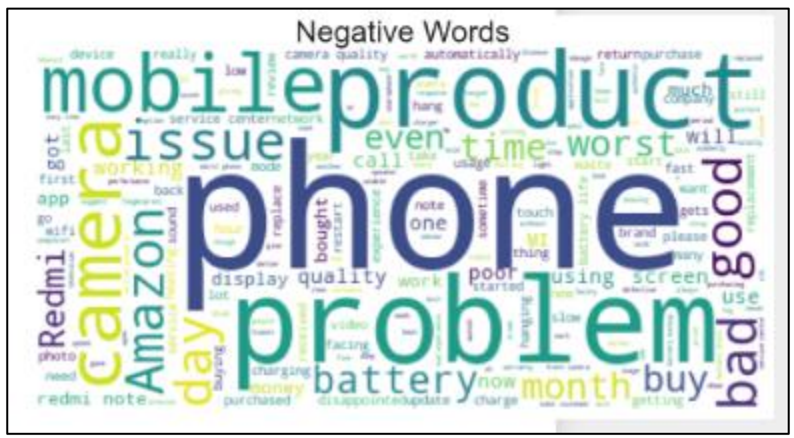
- Since we are classifying only the negative and positive sentiments we will consider only those and further using the LabelEncoder from sklearn.pre-processing we will convert the text categories to numeric categories as follows

data['y'] - Positives as 1 , Negatives as 0

	stars	reviews	Cleaned_reviews	tokens	POS_Tagging	Lemmas	sentiment_scores	Scores	compound	comp_score	y
0	1	The media could not be loaded.\n ...	media could loaded phone hanged many times ret...	[media, could, loaded, phone, hanged, many, ti...	[(media, n), (could, None), (loaded, v), (phon...	medium could load phone hang many time retur...	-0.42	{'neg': 0.158, 'neu': 0.766, 'pos': 0.077, 'co...	-0.4215	neg	0
1	5	Febulas performance Redmi Note 8 ...I love it ...	febulas performance redmi note 8 love first ti...	[febulas, performance, redmi, note, 8, love, f...	[(febulas, n), (performance, n), (redmi, v), (...	febulas performance redmi note 8 love first ...	0.91	{'neg': 0.0, 'neu': 0.451, 'pos': 0.549, 'comp...	0.9081	pos	1
2	5	best mobile under 10000	best mobile 10000	[best, mobile, 10000]	[(best, r), (mobile, a), (10000, None)]	best mobile 10000	0.64	{'neg': 0.0, 'neu': 0.323, 'pos': 0.677, 'comp...	0.6369	pos	1
3	5	Redmi note 8 is the best Smartphone under 10k ...	redmi note 8 best smartphone 10k year 2019	[redmi, note, 8, best, smartphone, 10k, year, ...	[(redmi, a), (note, n), (8, None), (best, a), ...	redmi note 8 best smartphone 10k year 2019	0.64	{'neg': 0.0, 'neu': 0.588, 'pos': 0.412, 'comp...	0.6369	pos	1
4	5	Loving the phone....Purchased with bank discou...	loving phone purchased bank discount 6gb 128gb...	[loving, phone, purchased, bank, discount, 6gb, 6gb...	[(loving, v), (phone, n), (purchased, v), (ban...	love phone purchase bank discount 6gb 128gb ...	0.99	{'neg': 0.0, 'neu': 0.659, 'pos': 0.341, 'comp...	0.9940	pos	1


All Words
phone
good
camera
mobile
battery life
battery
Redmi Note
quality
product
problem
great
day
issue
using
one
use
screen
time
buy
Overall
performance
display
camera quality
Amazon
better
excellent
charging
best
feature
price
working
month
device
much
look


Positive Words
phone
good
camera
great
quality
mobile
battery
battery life
price
Redmi
Note
day
product
performance
one
display
issue
camera quality
nice
excellent
best
10k
look
screen
buy
use
will
much
problem
work
better
feature
even
time
month
Amazon
charging
Overall
awesome


Negative Words
phone
mobile
product
problem
camera
issue
Amazon
good
bad
battery
worst
time
even
day
month
buy
Redmi
quality
display
poor
using
screen
one
bought
call
working
redmi note
purchased
disappointed
update
automatically
camera quality
service center
network

Feature Engineering

Using Bag of Words(BOW) and Tf-IDF term frequency-inverse document frequency

1. The bag-of-words (BOW) model converts text into fixed-length vectors by counting how many times each word appears

2. TF-IDF model contains information on the more important words and the less important ones as well.

```
# Bag of Words(BOW) Feature Extraction
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer()
text_counts= cv.fit_transform(data['cleaned_reviews'])
text_counts
```

```
<3099x8213 sparse matrix of type '<class 'numpy.int64'>'
	with 73739 stored elements in Compressed Sparse Row format>
```

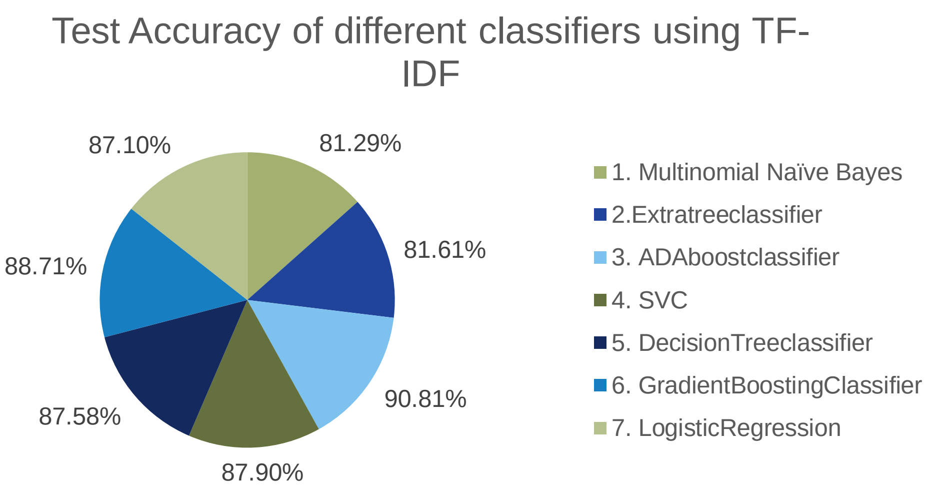
```
# TFIDF Feature Extraction
from sklearn.feature_extraction.text import TfidfVectorizer
tf=TfidfVectorizer()
tfidf_scores= tf.fit_transform(data['cleaned_reviews'])
tfidf_scores
```

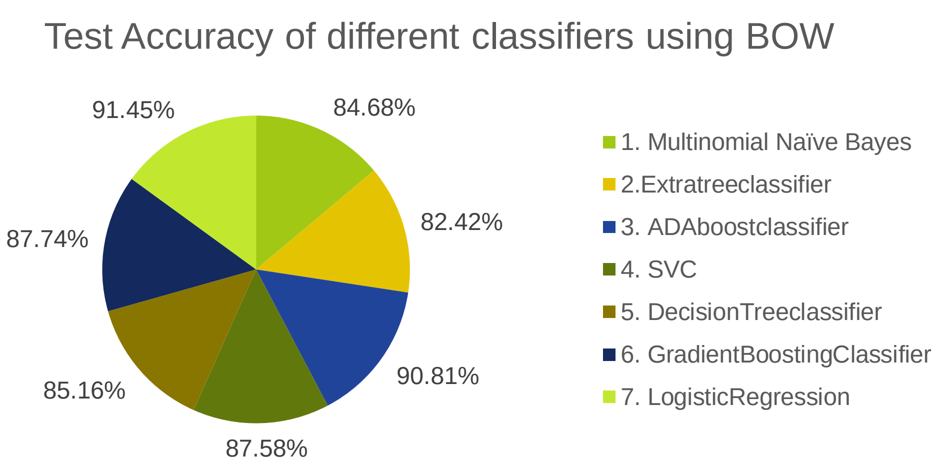
```
<3099x8213 sparse matrix of type '<class 'numpy.float64'>'
	with 73739 stored elements in Compressed Sparse Row format>
```

Model Building(Imbalanced Data)

Using both Bag of Words(BOW) and Tf-IDF term frequency-inverse document frequency we will build different classifier models for imbalanced data

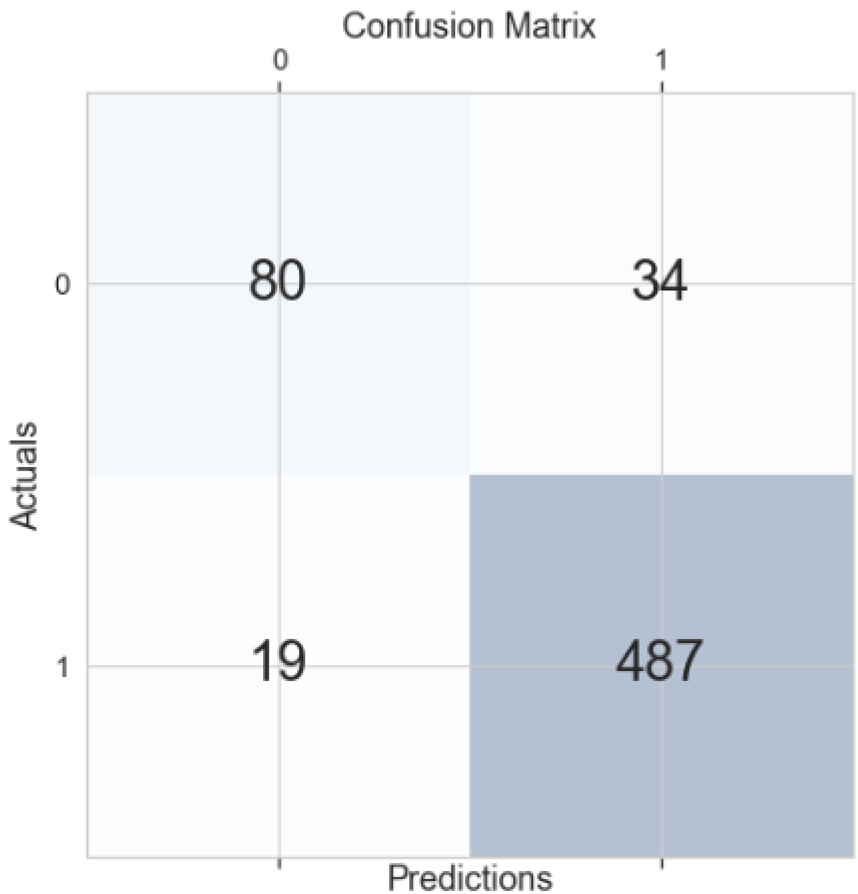
We have build 7 different models for both feature engineering techniques using train_test_split from sklearn.model_selection for classifying positive and negative sentiments and the accuracy results are as follows -





Models	TF-IDF		scores based on test predictions			BOW		scores based on test predictions		
	Train accuracy	Test accuracy	precision	recall	f1 score	Train accuracy	Test accuracy	precision	recall	f1 score
1. Multinomial Naïve Bayes	81.36%	81.29%	99.60%	81.55%	89.67%	90.76%	84.68%	88.33%	92.54%	90.39%
2.Extratreeclassifier	100.00%	81.61%	99.60%	81.81%	89.83%	100.00%	82.42%	99.60%	82.48%	90.24%
3. ADAboostclassifier	90.76%	90.81%	96.83%	92.27%	94.50%	90.48%	90.81%	95.65%	93.25%	94.43%
4. SVC	99.56%	87.90%	99.01%	87.74%	93.06%	95.16%	87.58%	98.02%	88.09%	92.79%
5. DecisionTreeclassifier	100.00%	87.58%	92.68%	92.14%	92.41%	100.00%	85.16%	90.90%	90.90%	90.90%
6.GradientBoostingClassifier	92.98%	88.71%	96.44%	90.37%	93.30%	92.25%	87.74%	96.83%	89.09%	92.80%
7. Logistic Regression	91.04%	87.10%	99.01%	86.97%	92.60%	99.48%	91.45%	96.24%	93.47%	94.83%

Thus the model with the highest accuracy is Logistic regression demonstrating an accuracy for testing of 91.45% with BOW as the feature extraction technique and the following are the statistics for same



Total count of predicted sentiments :
0s -99
1s – 521
Predictions total = 620

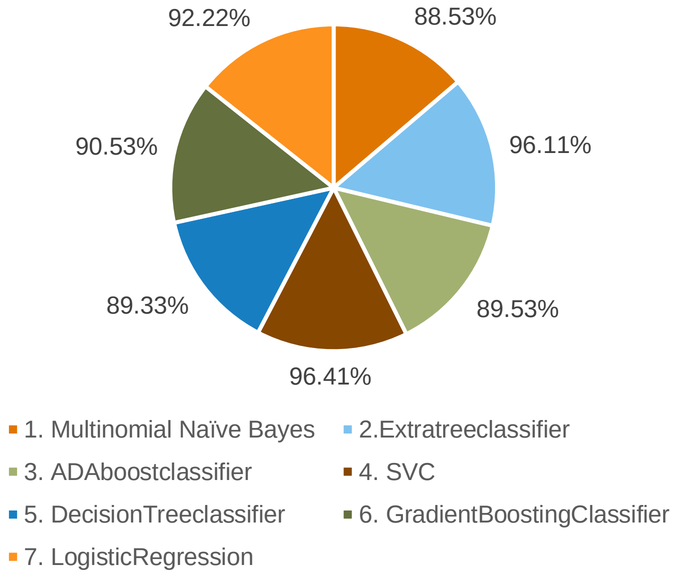
Actual sentiment counts in the testing dataset :
0s – 114
1s – 506
Actual total = 620

Model Building(Balanced Data)

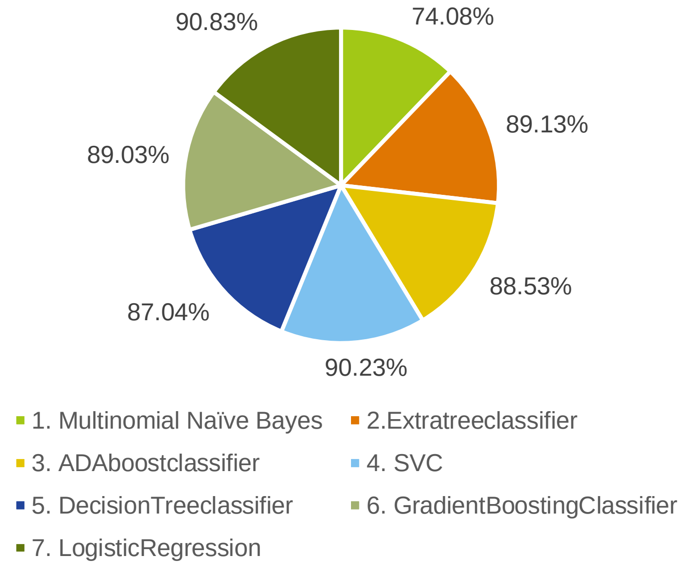
Using SMOTE to balance out the data for both Bag of Words(BOW) and Tf-IDF term frequency-inverse document frequency we will build different classifier models and compare to find the best model for further deployment

As in the previous slide we can see that the number of positive and negatives were unbalanced which was causing the biasness in the model. Therefore to improve the model SMOTE *(Oversampling technique to make the same count of negatives(0s) as positives(1s))* was used for both feature extraction techniques to balance the data by creating synthetic samples by doing upsampling. Further we have build 7 different models for both feature engineering techniques using train_test_split from sklearn.model_selection for classifying positive and negative sentiments and the accuracy results are as follows -

Test accuracy of different models using TFIDF

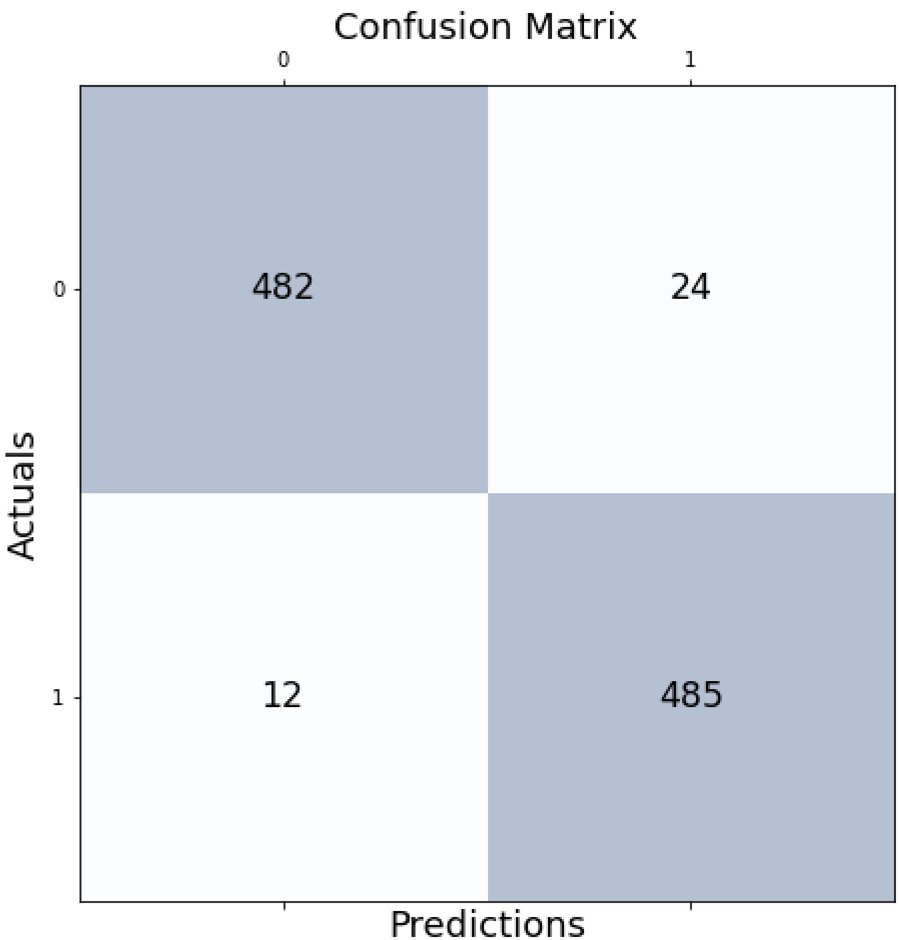


Test accuracy of different models using BOW



After Balancing the data using SMOTE										
Models	TF-IDF scores based on test predictions					BOW scores based on test predictions				
	Train accuracy	Test Accuracy	precision	recall	f1 score	Train accuracy	Test accuracy	precision	recall	f1 score
1. Multinomial Naïve Bayes	92.72%	88.53%	80.88%	95.26%	87.48%	79.18%	74.08%	85.31%	69.39%	76.53%
2.Extratreeclassifier	100.00%	96.11%	93.36%	98.72%	95.96%	99.83%	89.13%	91.14%	87.45%	89.26%
3. ADAboostclassifier	90.88%	89.53%	90.94%	89.59%	89.53%	89.55%	88.53%	84.41%	92.07%	87.90%
4. SVC	99.93%	96.41%	98.18%	94.75%	96.44%	96.71%	90.23%	91.46%	89.34%	90.23%
5. DecisionTreeclassifier	100.00%	89.33%	88.12%	90.12%	89.11%	99.83%	87.04%	86.51%	87.22%	86.86%
6. GradientBoostingClassifier	94.37%	90.53%	91.95%	89.25%	90.58%	90.88%	89.03%	84.70%	92.52%	88.44%
7. LogisticRegression	96.73%	92.22%	88.73%	95.24%	91.87%	98.23%	90.83%	88.53%	92.63%	90.53%

As a conclusion, the model having the highest accuracy, precision and recall scores will be our best fit for the deployment. So from the previous table we can see that the ExtraTreeClassifier and SVC both using TF-IDF are having highest values of all three metrics. But we will chose the SVC as the best model amongst all other models as it shows much better metric scores as compared to the ExtraTreeClassifier and the statistics for same can also be seen below –



Total count of predicted sentiments :
0s - 509
1s – 494
Predictions total = 1003

Actual sentiment counts in the testing dataset :
0s – 506
1s – 497
Actual total = 1003

Precision – 98.18%
Recall – 94.75%
F1-score – 96.44%

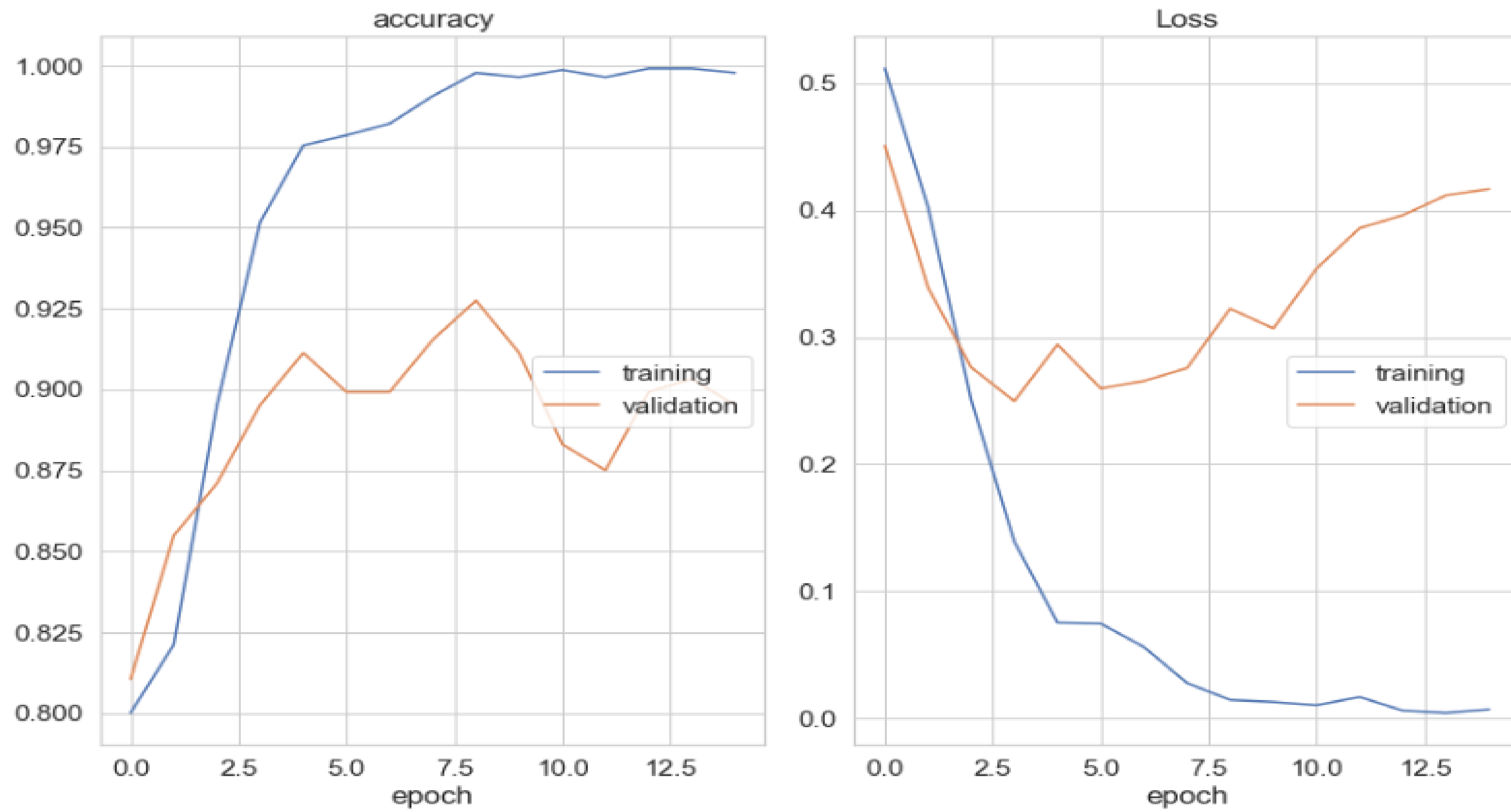
Deep Learning Model (Unbalanced Data)

Bi-directional LSTM Architecture using tensorflow.keras package

Bidirectional long-short term memory(bi-lstm) is the process of making any neural network o have the sequence information in both directions backwards (future to past) or forward(past to future). In bidirectional, our input flows in two directions, making a bi-lstm different from the regular LSTM.

Model Summary -

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 100, 64)           640000
_________________________________________________________________
spatial_dropout1d (SpatialDr (None, 100, 64)           0
_________________________________________________________________
bidirectional (Bidirectional (None, 256)               197632
_________________________________________________________________
dense (Dense)                (None, 1)                 257
=================================================================
Total params: 837,889
Trainable params: 837,889
Non-trainable params: 0
_________________________________________________________________
```




```
accuracy
	training           (min:    0.800, max:    0.999, cur:    0.998)
	validation         (min:    0.810, max:    0.927, cur:    0.895)
Loss
	training           (min:    0.004, max:    0.511, cur:    0.007)
	validation         (min:    0.250, max:    0.451, cur:    0.417)

 Loss:0.6593921780586243
 Accuracy:0.8629032373428345
              precision    recall  f1-score   support

           0       0.68      0.59      0.63       123
           1       0.90      0.93      0.92       497

    accuracy                           0.86       620
   macro avg       0.79      0.76      0.77       620
weighted avg       0.86      0.86      0.86       620
```

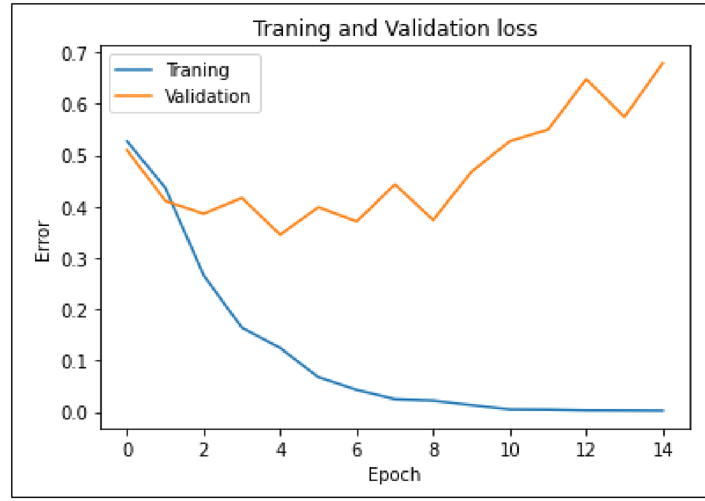
Deep Learning Model (Balanced Data)

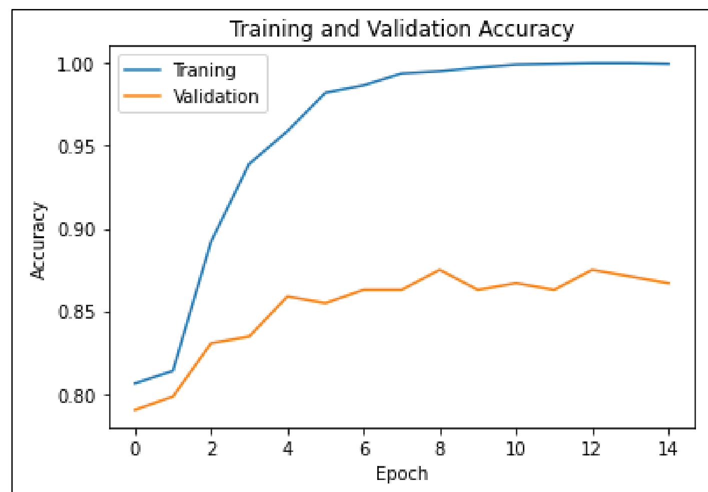
Balancing data by using EarlyStopping and building Bi-directional LSTM Architecture

Model Summary -

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 100, 64)           640000
_________________________________________________________________
spatial_dropout1d_1 (Spatial (None, 100, 64)           0
_________________________________________________________________
bidirectional_1 (Bidirection (None, 256)               197632
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 257
=================================================================
Total params: 837,889
Trainable params: 837,889
Non-trainable params: 0
_________________________________________________________________
```

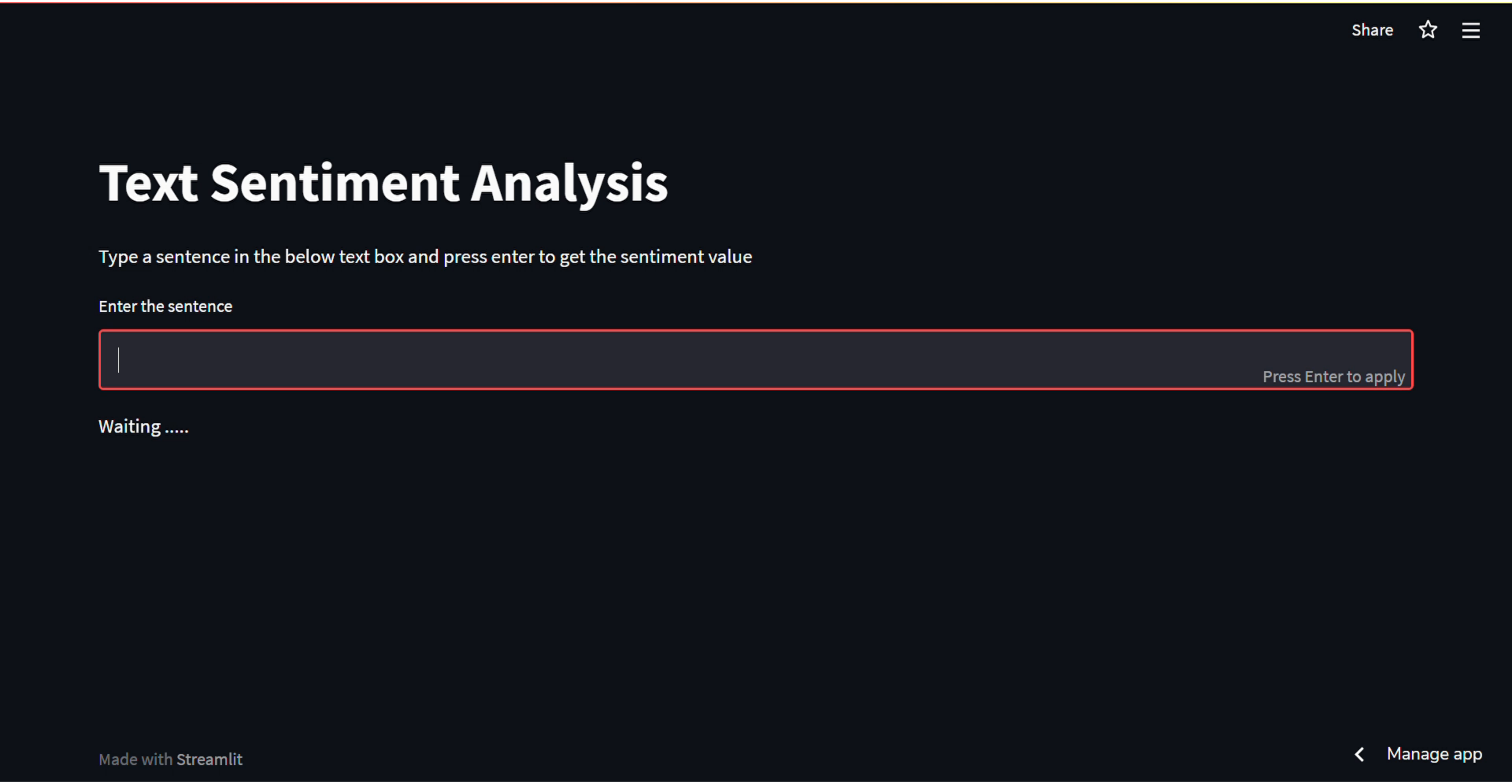
```
Loss:0.6091829538345337
Accuracy:0.8790322542190552
Precision: 0.918812
Recall: 0.931727
F1 score: 0.925224
              precision    recall  f1-score   support

           0       0.70      0.66      0.68       122
           1       0.92      0.93      0.93       498

    accuracy                           0.88       620
   macro avg       0.81      0.80      0.80       620
weighted avg       0.88      0.88      0.88       620
```





Deployment

Live app link - https://share.streamlit.io/triptideshpande/sentimentanalysis_nlp_project/main/Final_Deployment.py



Thank you!