VISVESVARAYA TECHNOLOGICAL UNIVERSITY

JNANA SANGAMA, BELAGAVI – 590 018

An Internship Project Report on

"AMAZON FOOD REVIEW SENTIMENT ANALYSIS"

Submitted in partial fulfillment of the requirements as a part of the

AI/ML INTERNSHIP

(NASTECH)

For the award of degree of

Bachelor of Engineering
in
Information Science and Engineering

Submitted by

PRAJVAL S	**PRADYUMNA DS**
1RN18IS076	**1RN18IS 075**

Internship Project Coordinators

Dr. R Rajkumar	**Mrs.Sunitha K**
Associate Professor	**Assistant Professor**
Dept. of ISE, RNSIT	**Dept. of ISE, RNSIT**

Department of Information Science and Engineering

RNS Institute of Technology

Channasandra, Dr. Vishnuvardhan Road, RR Nagar Post,
Bengaluru – 560 098

2021 -2022

RNS Institute of Technology

Channasandra, Dr. Vishnuvardhan Road, RR Nagar Post,
Bengaluru – 560 098

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the mini project report entitled ***AMAZON FOOD REVIEW SENTIMENT ANALYSIS*** has been successfully completed by **PRAJVAL S** bearing USN **1RN18IS076** and **PRADYUMNA DS** bearing USN **1RN18IS075** , presently VII semester students of **RNS Institute of Technology** in partial fulfillment of the requirements as a part of the ***AI/ML Internship (NASTECH)*** for the award of the degree of ***Bachelor of Engineering in Information Science and Engineering*** under **Visvesvaraya Technological University, Belagavi** during academic year **2021 – 2022**. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the report and deposited in the departmental library. The mini project report has been approved as it satisfies the academic requirements as a part of Mobile.

Dr. R Rajkumar	**Mrs. Sunitha K**	**Dr. Suresh L**
Coordinator	Guide	Professor and HoD
Associate Professor	Assistant Professor	

External Viva

Name of the Examiners **Signature with date**

1. ____________________________ ____________________________

2. ____________________________ ____________________________

ABSTRACT

Sentiment analysis refers to the task of natural language processing to determine whether a piece of text contains some subjective information and what subjective information it expresses, i.e., whether the attitude behind this text is positive, negative or neutral. Understanding the opinions behind user-generated content automatically is of great help for commercial and political use, among others. The task can be conducted on different levels, classifying the polarity of words, sentences or entire documents

Sentiment analysis and opinion mining have been acquiring a crucial role in both commercial and research applications because of their possible applicability to several different fields. Therefore a large number of companies have included the analysis of opinions and sentiments of customers as part of their mission. One of the most interesting applications of these approaches involves the automatic analysis of social network messages or customer reviews, on the basis of the feelings and emotions conveyed. This chapter aims to relate the most recent state-of-the-art sentiment-based techniques and tools to the affective characterization that may be inferred from social networks. The main result consists of a review of the most interesting methods employed to compare and classify customer reviews of Amazon food delivery and a description of advanced tools in this area.

To support sentiment analysis, various toolkits such as the Natural language toolkit (NLTK), open CV, Pattern, and SK Learn packages NLTK support preprocessing of text contents, and also offer the Naïve Bayes supervisor to implement frequency of terms analysis.

ACKNOWLEDGMENT

The fulfillment and rapture that go with the fruitful finishing of any assignment would be inadequate without the specifying the people who made it conceivable, whose steady direction and support delegated the endeavors with success.

We would like to profoundly thank **Management** of **RNS Institute of Technology** for providing such a healthy environment to carry out this AI/ML Internship Project.

We would like to express our thanks to our Principal **Dr. M K Venkatesha** for his support and inspired us towards the attainment of knowledge.

We wish to place on record our words of gratitude to **Dr. Suresh L,** Professor and Head of the Department, Information Science and Engineering, for being the enzyme and master mind behind our Mobile Application Development Laboratory with Mini Project Work.

We like to express our profound and cordial gratitude to my Internship Project Coordinators**, Dr. R Rajkumar,** Associate Professor, Department of Information Science and Engineering for their valuable guidance, constructive comments, continuous encouragement throughout the Mini Project Work and guidance in preparing report.

We would like to thank all other teaching and non-teaching staff of Information Science & Engineering who have directly or indirectly helped us to carry out the Mini Project Work.

Also, we would like to acknowledge and thank our parents who are source of inspiration and instrumental in carrying out this Mini Project Work.

PRAJVAL S	**PRADYUMNA DS**
USN:1RN18IS076	**USN:1RN18IS075**

TABLE OF CONTENTS

Abstract **iii**

Acknowledgement **iv**

Table of Content **v**

List of Figures **vi**

1. **INTRODUCTION** 01

1.1. ORGANIZATION/ INDUSTRY 01

1.1.1. Company Profile 01

1.1.2. Domain/ Technology (Data Science/Mobile computing/…) 01

1.1.3. Department/ Division / Group 02

1.2. PROBLEM STATEMENT 02

1.2.1. Existing System and their Limitations 02

1.2.2. Proposed Solution 02

1.2.3. Problem formulation 02

2. **REQUIREMENT ANALYSIS, TOOLS &TECHNOLOGIES** 03

2.1. Hardware & Software Requirements 03

2.2. Tools/ Languages/ Platform 03

3. **DESIGN AND IMPLIMENTATION** 03

3.1. Architecture/ DFD/Sequence diagram/Class diagrams /Flowchart 03

3.2. Problem statement 04

3.3. Algorithm/Methods/ Pseudo code 06

3.4. Libraries used / API’S 06

4. **OBSERVATIONS AND RESULTS** 9

4.1. Testing 9

4.2. Results & Snapshots 10

5. **CONCLUSION AND FUTURE WORK** 13

5.1. Conclusion 13

5.2. Future Enhancement 14

6. **REFERENCES**

LIST OF FIGURES

3.1 DESCRIPTION OF DATASET

3.2 WORD CLOUD BEFORE PROCESSING

3.3 WORD CLOUD AFTER TEXT PROCESSING

3.4 LOGISTIC REGRESSION ALGORITHM

3.5 NAÏVE BAYES ALGORITHM

4.1 CONFUSION MATRIX OF LOGISTIC REGRESSION

4.2 CONFUSION MATRIX FOR NAÏVE BAYES ALGORITHM

4.3 SETIMENT PREDICTION OUTPUT

Chapter 1

INTRODUCTION

1.1 ORGANIZATION/INDUSTRY

1.1.1 COMPANY PROFILE

NASTECH is formed with the purpose of bridging the gap between Academia and Industry Nastech is one of the leading Global Certification and Training service providers for technical and management programs for educational institutions. We collaborate with educational institutes to understand their requirements and form a strategy in consultation with all stakeholders to fulfill those by skilling, reskilling and upskilling the students and faculties on new age skills and technologies.

1.1.2 DOMAIN/TECHNOLOGY

The domain chosen for our project is AI/ML. Machine learning, the fundamental driver of AI, is possible through algorithms that can learn themselves from data and identify patterns to make predictions and achieve your predefined goals, rather than blindly following detailed programmed instructions, like in traditional computer programming. This technology allows the machine to perceive, learn, reason and communicate through observation of data, like a child that grows up and acquires knowledge from examples. Machines also have the advantage of not being limited by our inherent biological limitations. With machine learning, manufacturing companies have increased production capacity up to 20%, while lowering material consumption rates by 4%.

Nowadays, the revolutionary AI technology evolved from rule-based expert systems to machine learning and more advanced subcomponents such as deep learning (learning representations instead of tasks), artificial neural networks (inspired by animal brains) and reinforcement learning (virtual agents rewarded if they made good decisions).

The AI can master the complexity of the intertwining industrial processes to enhance the whole flow of production instead of isolated processes. This enormous cognitive capacity gives the AI the ability to consider the spatial organization of plants and the timing constraints of live production. Another key advantage is the capability of AI algorithms to think probabilistically, with all the subtlety this allows in edge cases, instead of traditional rulebased methods that require rigid theories and a full comprehension of problems.

1.1.3 Department

R.N.Shetty Institute of Technology (RNSIT) established in the year 2001, is the brain-child of the Group Chairman, Dr. R. N. Shetty. The Murudeshwar Group of Companies headed by Sri. R. N. Shetty is a leading player in many industries viz construction, manufacturing, hotel, automobile, power & IT services and education. The group has contributed significantly to the field of education. A number of educational institutions are run by the

R. N. Shetty Trust, RNSIT being one amongst them. With a continuous desire to provide quality education to the society, the group has established RNSIT, an institution to nourish and produce the best of engineering talents in the country. RNSIT is one of the best and top accredited engineering colleges in Bengaluru.

1.2 PROBLEM STATEMENT

1.2.1 Existing System and their Limitations

A manual method is currently used in the market to analyse the sentiment of the customers. This method is both time consuming and inefficient. In order to analyse the sentiment of the customers on how the products are received by them we need a better model which is efficient and less time consuming.

1.2.2 Proposed Solution

To eliminate the drawbacks stated above, Natural Language Processing packages coupled with machine learning algorithms can be used to implement the model which helps in sentiment analysis of Amazon food review system.

1.2.3 Program formulation

Sentiment analysis is the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and Subjective information

Chapter 2

REQUIREMENT ANALYSIS, TOOLS &TECHNOLOGIES

2.1 Hardware and Software Requirements

2.1.1 Hardware Requirements:

- Processor: Pentium IV or above
- RAM: 4 GB or more
- Hard Disk: 2GB or more

2.1.2 Software Requirements:

- Operating System: Windows 7 or above
- IDE: Google Colab

2.2 Tools/Languages/Platforms

- Python

CHAPTER 3

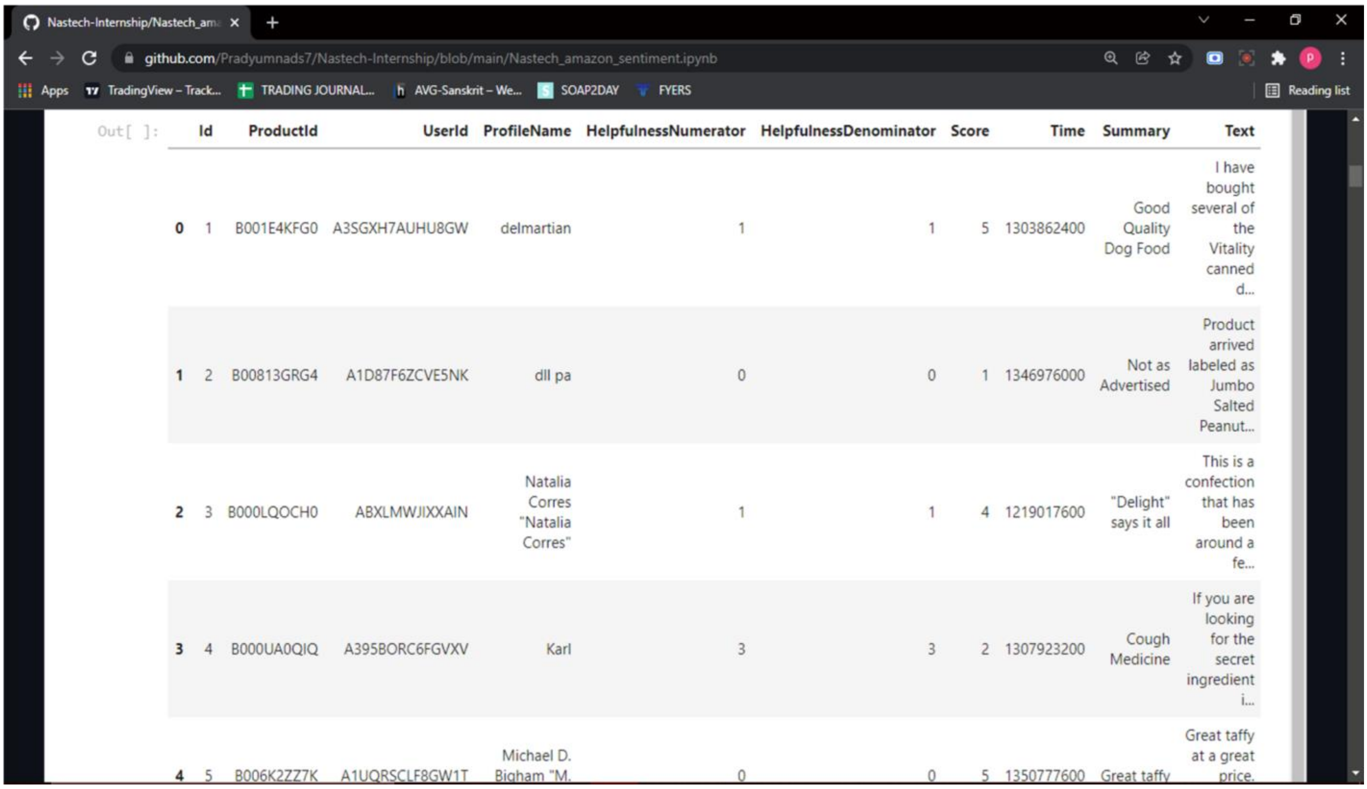
3.1 Problem Statement

The goal of this statistical analysis is to help us understand the sentiment behind the reviews and how these reviews can be used to improve the services. Our objective is to predict the sentiment behind the reviews given by customers.

NLP models and various machine learning tools have been used to obtain the desired result

The following features have been used:

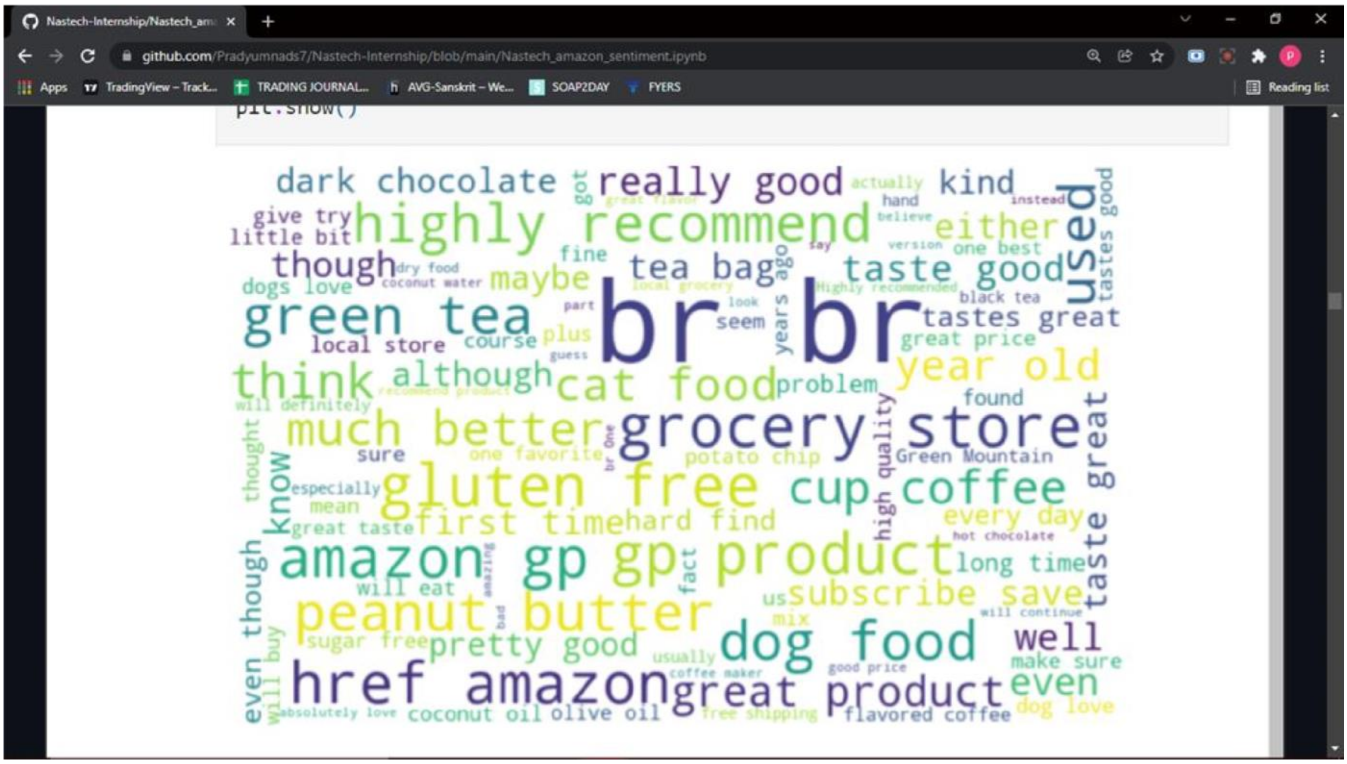
1. Product Id : Unique identifier for the product
2. User Id : Unique identifier for the user
3. Profile Name : Profile name of the user
4. Helpfulness Numerator : Number of users who found the review helpful
5. Helpfulness Denominator : Number of users who indicated whether they found the review helpful or not.
6. Score : Rating between 1 and 5
7. Timestamp : Time
8. Summary : Summary of the review
9. Text : The actual review


Nastech-Internship/Nastech_am ×
github.com/Pradyumnads7/Nastech-Internship/blob/main/Nastech_amazon_sentiment.ipynb
Apps
TradingView – Track...
TRADING JOURNAL...
AVG-Sanskrit – We...
SOAP2DAY
FYERS
Reading list

Out[ ]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M.	0	0	5	1350777600	Great taffy	Great taffy at a great price.

Figure 3.1 Description of dataset



```
plt.show()
```

Figure 3.2 Word cloud before processing

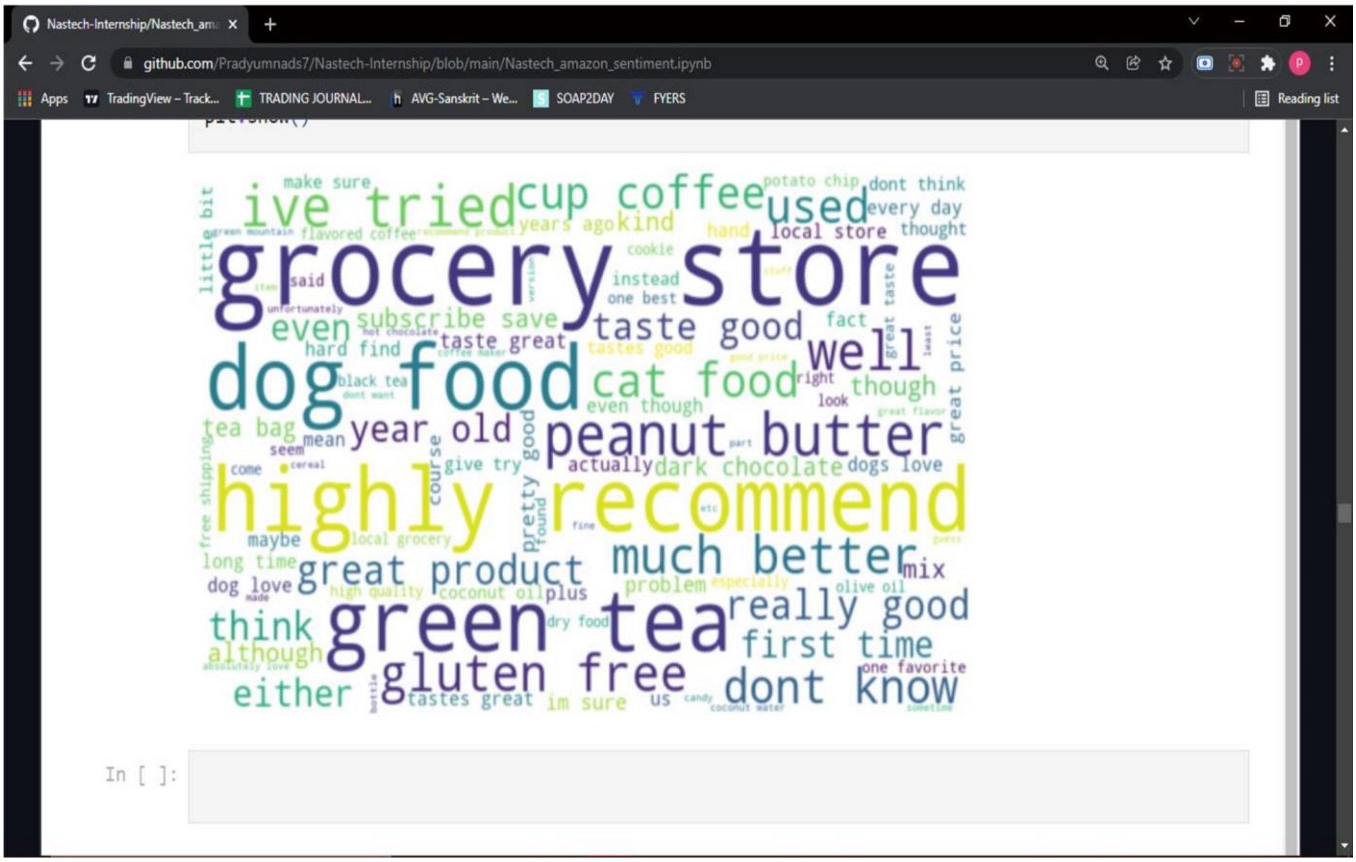


Figure 3.3 Word cloud post processing

The figure 3.2 gives us the view of word cloud of data before it is processed Many stop words like “will”, “us”,”and” , “href” have multiple occurrence as shown above in the word cloud

The figure 3.3 gives us the view of word cloud after it is processed and the un wanted stop words are removed for better understanding of the review

Word Cloud or Tag Clouds is a visualization technique for texts that are natively used for visualizing the tags or keywords from the websites

3.2 Algorithm

Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False.

The sigmoid function is a mathematical function used to map the predicted values to probabilities.

It maps any real value into another value within a range of 0 and 1
The formula for logistic regression is as follows

$$\boldsymbol{log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n}$$

Figure 3.4 Logistic regression formula

Naïve Bayes

Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.

The formula is given as

$$\mathrm{P(A|B)}= \frac{P(B|A)P(A)}{P(B)}$$

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.
P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

3.3 Libraries

- Pandas
- Numpy
- Seaborn
- Matplotlib
- Sklearn
- NLTK

3.4 Data Processing

1) Cleaning the review text in the dataset so that the text can be converted to a format which can be processed by the machine learning algorithms.

```
def clean_text(Review):

    Review = str(Review).lower() # convert to lowercase
    Review = re.sub('\[.*?\]', '', Review)
    Review = re.sub('https?://\S+|www\.\S+', '', Review) # Remove URls
    Review = re.sub('<.*?>+', '', Review)
    Review = re.sub(r'[^a-z0-9\s]', '', Review) # Remove punctuation
    Review = re.sub('\n', '', Review)
    Review = re.sub('\w*\d\w*', '', Review)
    return Review
```

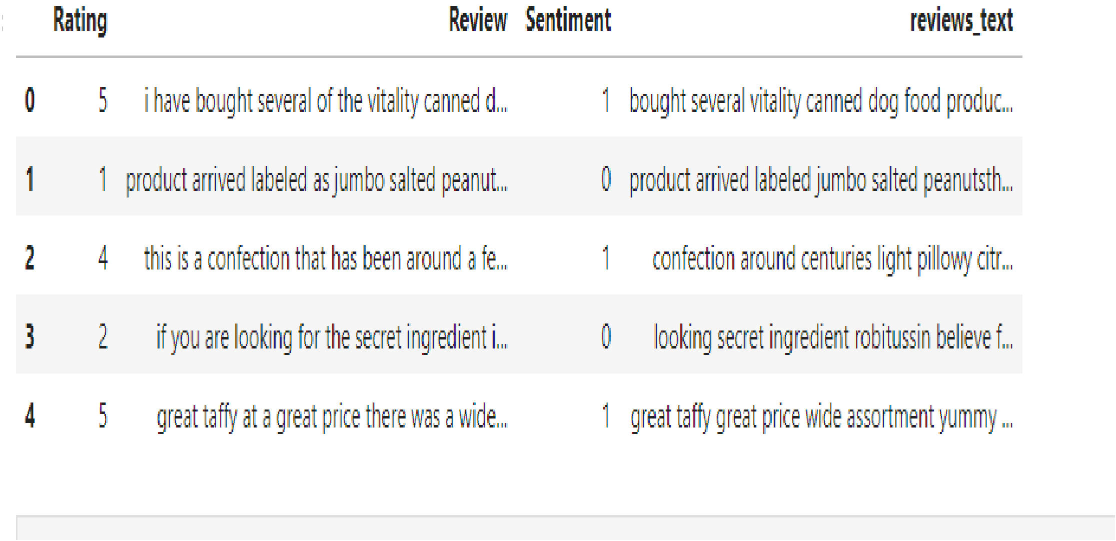
2) Removal of Stop words from the text as they are not helpful when processing the text for sentiment analysis.

```
import nltk
nltk.download('stopwords')
nltk.download('punkt')
from nltk.corpus import stopwords

stop_words = set(stopwords.words('english'))
stopword = []
sentence = df['Review'][0]


#words = nltk.word_tokenize(sentence)
```

```
def remove_stopword(stop_words, sentence):
    return [word for word in nltk.word_tokenize(sentence) if word not in
stop_words]

df['reviews_text'] = df['Review'].apply(lambda row: '
'.join(remove_stopword(stop_words, row)))
```

3) Converting the given ratings to sentiment by considering ratings <=3 as 0 or Negative and ratings >3 as 1 or Positive.

```
def apply_sentiment(Rating):
    if(Rating <=3 ):
        return 0
    else:
        return 1
```



	Rating	Review	Sentiment	reviews_text
0	5	i have bought several of the vitality canned d...	1	bought several vitality canned dog food produc...
1	1	product arrived labeled as jumbo salted peanut...	0	product arrived labeled jumbo salted peanutsth...
2	4	this is a confection that has been around a fe...	1	confection around centuries light pillowy citr...
3	2	if you are looking for the secret ingredient i...	0	looking secret ingredient robitussin believe f...
4	5	great taffy at a great price there was a wide...	1	great taffy great price wide assortment yummy ...

Figure 3.5 Dataset after Data Processing

Chapter 4

OBSERVATION AND RESULTS

4.1 Training and Testing

- **Train Test split**

The Dataset has been split for training and testing by considering 80% data for training and 20% data for testing.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,random_state =
42,
                                                    test_size = 0.20)
```

- **Training Naïve Bayes model**

```
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import CountVectorizer,
TfidfVectorizer, TfidfTransformer
from sklearn.naive_bayes import MultinomialNB

clf = Pipeline([
    ('vect', CountVectorizer(stop_words='english')),
    ('tfidf', TfidfTransformer()),
    ('classifier', MultinomialNB()),
    ])
```

```
nb_model = clf.fit(X_train,y_train)
```

- **Training Logistic Regression model**

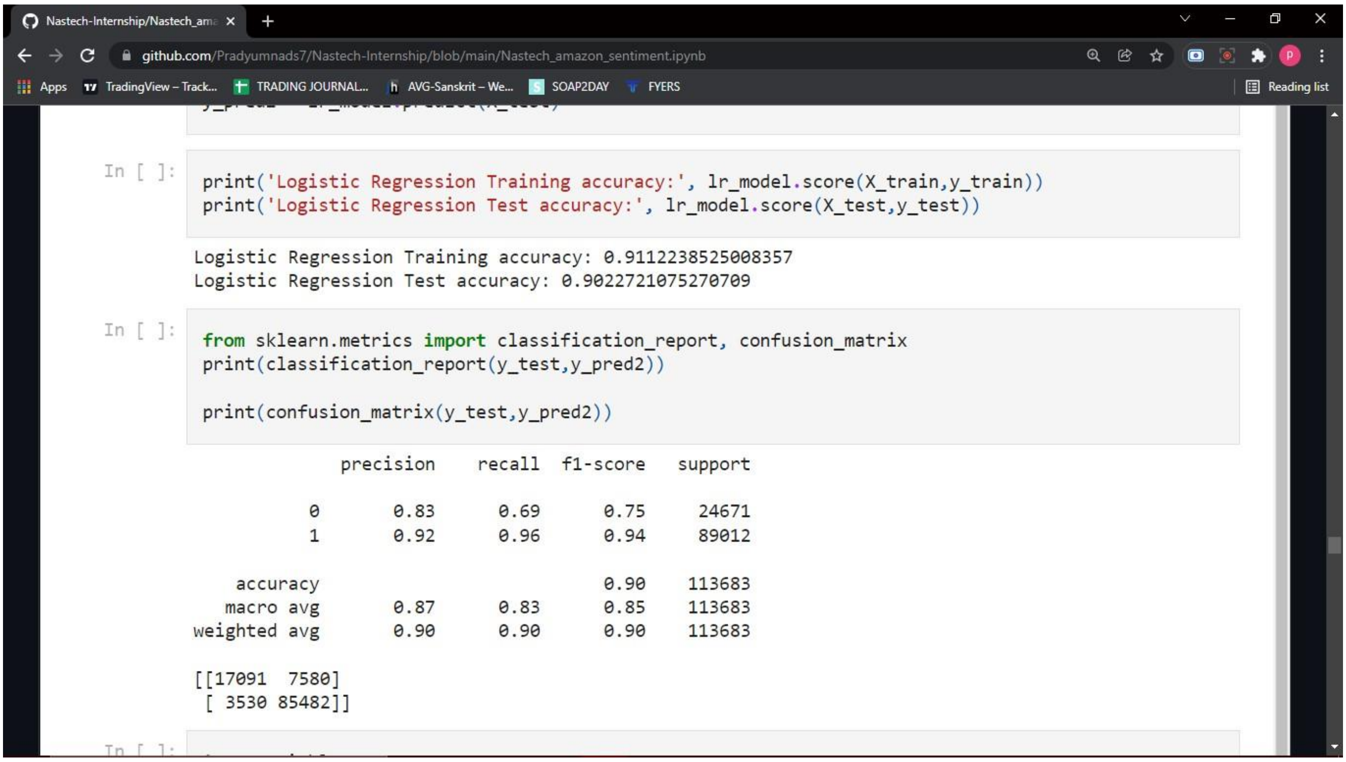
```
from sklearn.linear_model import LogisticRegression


clf2 = Pipeline([
    ('vect', CountVectorizer()),
    ('tfidf', TfidfTransformer()),
    ('classifier', LogisticRegression()),
    ])


lr_model = clf2.fit(X_train,y_train)
```

4.2 Results & Snapshots

Evaluation of Logistic Regression model

The Performance of the Logistic Regression model can be measured using the evaluation metrics such as Accuracy, Precision, F1 score. The model's Training and Testing accuracy is displayed as well.

```
y_pred2 = lr_model.predict(X_test)

print('Logistic Regression Training accuracy:',
lr_model.score(X_train,y_train))
print('Logistic Regression Test accuracy:',
lr_model.score(X_test,y_test))
```


Nastech-Internship/Nastech_ama
github.com/Pradyumnads7/Nastech-Internship/blob/main/Nastech_amazon_sentiment.ipynb
Apps
TradingView – Track...
TRADING JOURNAL...
AVG-Sanskrit – We...
SOAP2DAY
FYERS
Reading list
In [ ]:
In [ ]:

```
print('Logistic Regression Training accuracy:', lr_model.score(X_train,y_train))
print('Logistic Regression Test accuracy:', lr_model.score(X_test,y_test))
```

```
Logistic Regression Training accuracy: 0.9112238525008357
Logistic Regression Test accuracy: 0.9022721075270709
```

```
from sklearn.metrics import classification_report, confusion_matrix
print(classification_report(y_test,y_pred2))

print(confusion_matrix(y_test,y_pred2))
```

```
              precision    recall  f1-score   support

           0       0.83      0.69      0.75     24671
           1       0.92      0.96      0.94     89012

    accuracy                           0.90    113683
   macro avg       0.87      0.83      0.85    113683
weighted avg       0.90      0.90      0.90    113683

[[17091  7580]
 [ 3530 85482]]
```

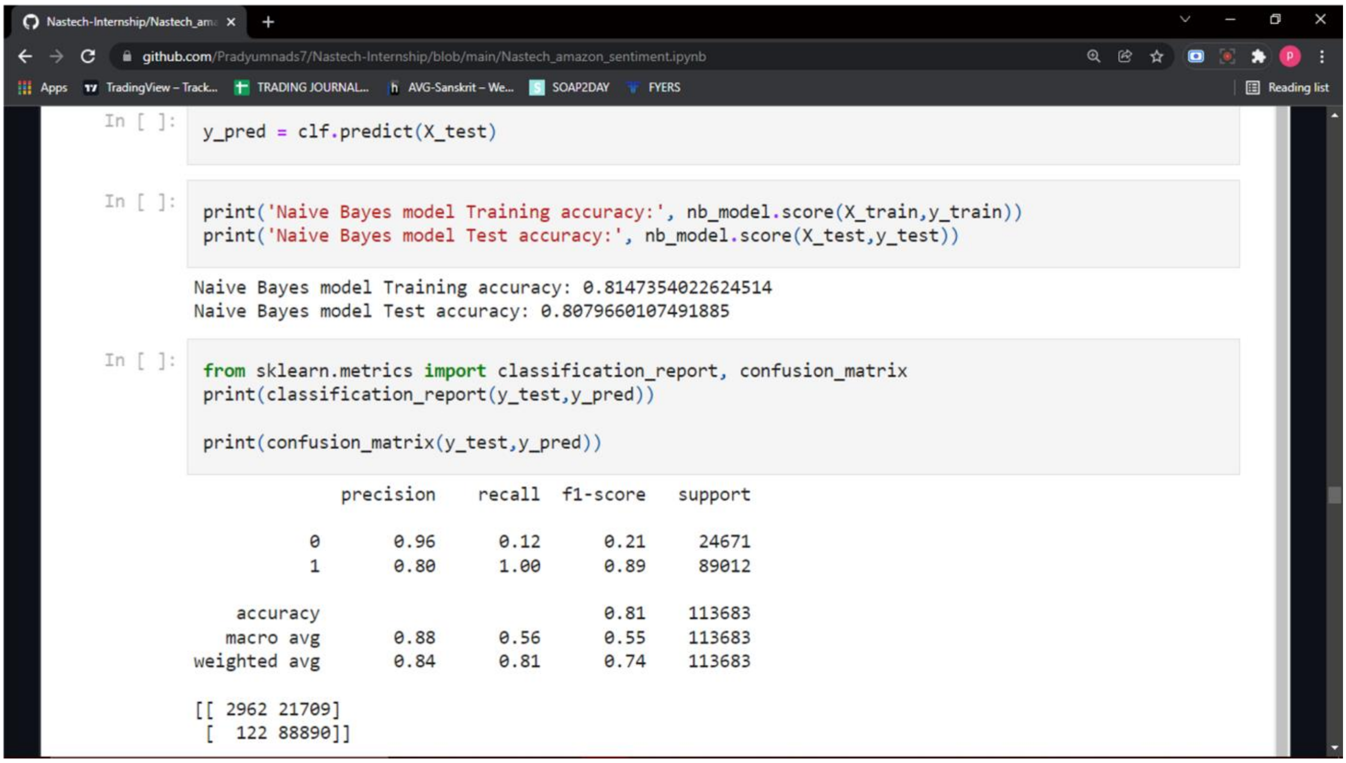
Figure 4.1 Classification Report for Logistic Regression

In the above fig 4.1, the f1 – score and the accuracy of the model is being shown using the confusion matrix

Evaluation of Naïve Bayes model

The Performance of the Naïve Bayes model can be measured using the evaluation metrics such as Accuracy, Precision, F1 score. The model's Training and Testing accuracy is displayed as well.

```
y_pred = clf.predict(X_test)


print('Naive Bayes model Training accuracy:',
nb_model.score(X_train,y_train))
print('Naive Bayes model Test accuracy:', nb_model.score(X_test,y_test))
```


Nastech-Internship/Nastech_ama
github.com/Pradyumnads7/Nastech-Internship/blob/main/Nastech_amazon_sentiment.ipynb
Apps
TradingView – Track...
TRADING JOURNAL...
AVG-Sanskrit – We...
SOAP2DAY
FYERS
Reading list

```
In [ ]:  y_pred = clf.predict(X_test)

In [ ]:  print('Naive Bayes model Training accuracy:', nb_model.score(X_train,y_train))
         print('Naive Bayes model Test accuracy:', nb_model.score(X_test,y_test))

        Naive Bayes model Training accuracy: 0.8147354022624514
        Naive Bayes model Test accuracy: 0.8079660107491885

In [ ]:  from sklearn.metrics import classification_report, confusion_matrix
         print(classification_report(y_test,y_pred))

         print(confusion_matrix(y_test,y_pred))

                      precision    recall  f1-score   support

                   0       0.96      0.12      0.21     24671
                   1       0.80      1.00      0.89     89012

            accuracy                           0.81    113683
           macro avg       0.88      0.56      0.55    113683
        weighted avg       0.84      0.81      0.74    113683

        [[ 2962 21709]
         [  122 88890]]
```

Figure 4.1 Classification Report for Naïve Bayes

Precision is a metric that quantifies the number of correct positive predictions made.

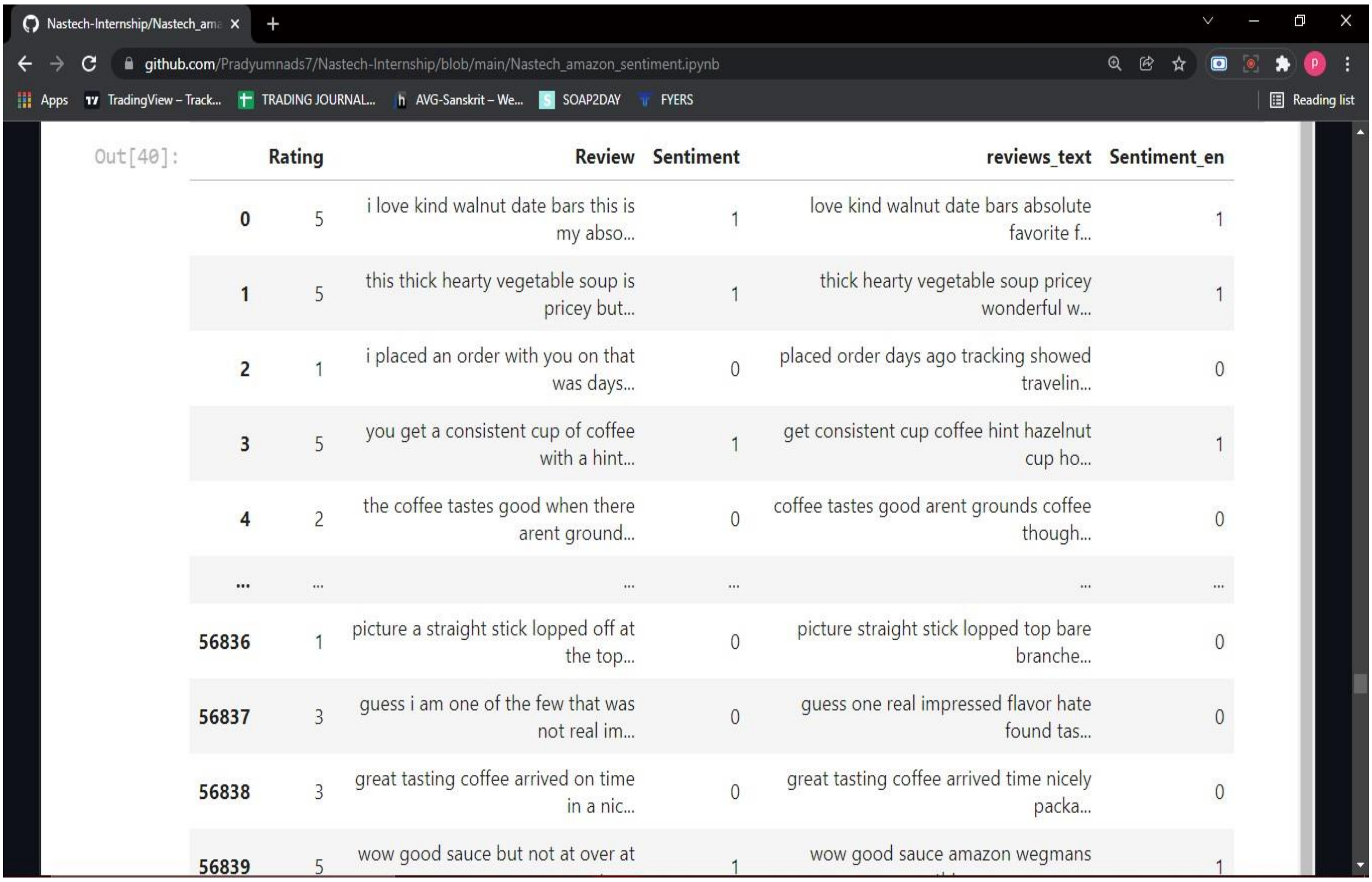
It is calculated as the ratio of correctly predicted positive examples divided by the total number of positive examples that were predicted.

Recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made.

Unlike precision that only comments on the correct positive predictions out of all positive predictions, recall provides an indication of missed positive predictions.

The F1 score is defined as the harmonic mean of precision and recall. ... Since the F1 score is an average of Precision and Recall, it means that the F1 score gives equal weight to Precision and Recall:

A model will obtain a high F1 score if both Precision and Recall are high


Nastech-Internship/Nastech_am
github.com/Pradyumnads7/Nastech-Internship/blob/main/Nastech_amazon_sentiment.ipynb
Apps
TradingView – Track...
TRADING JOURNAL...
AVG-Sanskrit – We...
SOAP2DAY
FYERS
Reading list

Out[40]:

	Rating	Review	Sentiment	reviews_text	Sentiment_en
0	5	i love kind walnut date bars this is my abso...	1	love kind walnut date bars absolute favorite f...	1
1	5	this thick hearty vegetable soup is pricey but...	1	thick hearty vegetable soup pricey wonderful w...	1
2	1	i placed an order with you on that was days...	0	placed order days ago tracking showed travelin...	0
3	5	you get a consistent cup of coffee with a hint...	1	get consistent cup coffee hint hazelnut cup ho...	1
4	2	the coffee tastes good when there arent ground...	0	coffee tastes good arent grounds coffee though...	0
...	...	...	...	...	...
56836	1	picture a straight stick lopped off at the top...	0	picture straight stick lopped top bare branche...	0
56837	3	guess i am one of the few that was not real im...	0	guess one real impressed flavor hate found tas...	0
56838	3	great tasting coffee arrived on time in a nic...	0	great tasting coffee arrived time nicely packa...	0
56839	5	wow good sauce but not at over at	1	wow good sauce amazon wegmans	1

Figure 4.3 Sentiment analysis output

The above figure gives us the sentiment of each review by using 0’s and 1’s as the analyzing metric. 1being a positive sentiment of rating > 3 and 0 being a metric for negative sentiment having a rating of less than or equal to 3.

Chapter 5

CONCLUSION AND FUTURE ENHANCEMENT

5.1 Conclusion

Aim of the project is to predict the sentiment behind the reviews of the customers taking into consideration various features such as ratings, emotions , helpfulness numerator , helpfulness denominator , summary etc. which has successfully been achieved with an accuracy of 91%. The proposed model is definitely the best substitute for the manual method which is inaccurate, less efficient and time consuming. Based on the results, it can be concluded that such ML-driven predictions are easily comprehendible and significant from a data-analytics point of view. When correctly implemented, a high rate of accuracy can be achieved.

5.2 Future Enhancement

To make the system even more efficient and user-friendly, Deep Learning tools can be included. This will enhance the overall efficiency of the sentimental analysis of the reviews with a better accuracy rate and f1 score. This can in turn help with better understanding of the sentiment behind the reviews of the customers. Various other machine learning algorithms can also be used apart from Naïve Bayes to improve the accuracy of the model.

Chapter 6

REFERENCE

1. McCallum, A., & Nigam, K. (1998). A comparison of event models for Naïve Bayes text classification. In AAAI-98 Workshop on Learning for Text Categorization (pp. 41–48)

2. An Introduction to Generalized Linear Models, by Dobson

3. Natural Language Processing with Python by Steven Bird , Ewan Klain and Edward Loper

4. Quora questions and answers

5. Stack overflow for resolving errors

6. Other external links