

# **SENTIMENTAL ANALYSIS ON AMAZON FINE FOOD REVIEW**

## **Mini Project Report**

Submitted by

**Rivin Cheria**

*Submitted in partial fulfillment of the requirements for the award of  
the degree of*

*Master of Computer Applications  
Of*

*A P J Abdul Kalam Technological University*



**FEDERAL INSTITUTE OF SCIENCE AND TECHNOLOGY (FISAT)®**

**ANGAMALY-683577, ERNAKULAM(DIST)**

**MARCH 2022**

## **DECLARATION**

I, **Rivin Cheria**, hereby declare that the report of this project work, submitted to the Department of Computer Applications, Federal Institute of Science and Technology (**FISAT**), Angamaly in partial fulfillment of the award of the degree of Master of Computer Application is an authentic record of our original work.

The report has not been submitted for the award of any degree of this university or any other university.

**Date : 04-03-2022**

**Place: Angamaly**

**FEDERAL INSTITUTE OF SCIENCE AND  
TECHNOLOGY (FISAT)®  
ANGAMALY, ERNAKULAM-683577**

**DEPARTMENT OF COMPUTER APPLICATIONS**



**CERTIFICATE**

This is to certify that the project report titled "**SENTIMENTAL ANALYSIS ON AMAZON FINE FOOD REVIEW**" submitted by **Rivin Cheria** towards partial fulfillment of the requirements for the award of the degree of Master of Computer Applications is a record of bonafide work carried out by them during the year 2022.

**Project Guide**

**Head of the Department**

Submitted for the viva-voice held on ..... at .....

**Examiner1 :**

**Examiner2 :**

## ACKNOWLEDGEMENT

Gratitude is a feeling which is more eloquent than words, more silent than silence. To complete this project work I needed the direction, assistance and co-operation of various individuals, which is received in abundance with the grace of God.

I hereby express our deep sense of gratitude to **Dr. Manoge George**, Principal of FISAT and **Dr. C Sheela**, Vice principal of FISAT, for allowing us to utilize all the facilities of the college.

My sincere thanks to **Dr. Deepa Mary Mathew**, Head of the department of Computer Applications FISAT and scrum master and our Internal guide for this project **Ms.Shidha and Ms.Joice T** for giving valuable guidance, constructive suggestions and comment during my project work. I also express my boundless gratitude to all the lab faculty members for their guidance.

Finally I wish to express a whole heart-ed thanks to my parents, friends and well-wishers who extended their help in one way or other in preparation of my project. Besides all, I thank GOD for everything.

## **ABSTRACT**

Social media has given ample opportunity to the consumer in terms of gauging the quality of the products by reading and examining the reviews posted by the users of online shopping platforms. Moreover, online platforms such as Amazon.com provides an option to the users to label a review as 'Helpful' if they find the content of the review valuable. This helps both consumers and manufacturers to evaluate general preferences in an efficient manner by focusing mainly on the selected helpful reviews. However, the recently posted reviews get comparatively fewer votes and the higher voted reviews get into the users' radars first. This study deals with these issues by building an automated text classification system to predict the helpfulness of online reviews irrespective of the time they are posted. The study is conducted on the data collected from Amazon.com consisting of the reviews on fine food. The focus of previous research has mostly remained on finding a correlation between the review helpfulness measure and review content-based features. In addition to finding significant content-based features, this study uses three different approaches to predict the review helpfulness which includes vectorized features, review and summary centric features, and word embedding-based features. Here i am using Multinomial Naive Bayes for implementing the model, which has an accuracy about 85percentage

# Contents

<b>1</b>	<b>INTRODUCTION</b>	<b>8</b>
<b>2</b>	<b>PROOF OF CONCEPT</b>	<b>9</b>
2.1	Existing System . . . . .	9
2.2	Proposed System . . . . .	9
2.3	Objectives . . . . .	10
<b>3</b>	<b>SCRUM MEETINGS</b>	<b>11</b>
<b>4</b>	<b>IMPLEMENTATION</b>	<b>14</b>
4.1	System Architecture . . . . .	16
4.2	Dataset . . . . .	16
4.3	Modules . . . . .	16
4.3.1	Data Preprocessing . . . . .	16
4.3.2	TEXT ENGINEERING . . . . .	17
<b>5</b>	<b>RESULT ANALYSIS</b>	<b>19</b>
<b>6</b>	<b>CONCLUSION AND FUTURE SCOPE</b>	<b>20</b>
6.1	Conclusion . . . . .	20
6.2	Future Scope . . . . .	21
<b>7</b>	<b>SOURCE CODE</b>	<b>22</b>

<b>8</b>	<b>SCREEN SHOTS</b>	<b>25</b>
<b>9</b>	<b>REFERENCES</b>	<b>28</b>

# Chapter 1

## INTRODUCTION

Nowadays, in the digital world users are very aware and particular about expressing their opinions on certain products once they have used, they love to write it as reviews in the ecommerce websites in the most lucid way. This has helped the concerned company identifying the issues for those products which they can able to fix faster as well and moreover, manufactures also become aware about the performance of those products in the market. Sentiment Analysis plays an important role in terms of detecting the contextual polarity of the reviews, in other way it is also known as opinion mining as it determines the attitude of the users towards the products from the huge volume of opinions i.e. reviews. Here, we have adopted certain preprocessing techniques as a part of the NLP such as tokenization, stemming or Lemmatization followed by the removal of stop-words, HTML tags with the help of different Python libraries. Then we have availed some of the vectorizer techniques such as Bag of Words, Tf-Idf which has converted texts into vectors so that the Machine Learning algorithms can be applied onto these. After converting into vectorized data, we have applied algorithm multinomial naive bayes is used to prognosticate the polarity of the reviews. We also have ratings of these products, if the ratings are 1 or 2 the corresponding products are said to fall under the negative polarity and if the ratings are 4 or 5 it belongs to the positive polarity. For those products having rated as 3, We have kept those as neutral.



## **Chapter 2**

# **PROOF OF CONCEPT**

### **2.1 Existing System**

In this study, the focus was on building an automated text classification system which can predict the helpfulness measure of an online review irrespective of the time of posting. The purpose of this was to provide both consumers and manufacturers a wide variety of reviews to choose from by including the most recent yet unvoted reviews in addition to higher voted old dated reviews.

### **2.2 Proposed System**

In this study, the focus was on building an automated text classification system which can predict the helpfulness measure of an online review irrespective of the time of posting. The purpose of this was to provide both consumers and manufacturers a wide variety of reviews to choose from by including the most recent yet unvoted reviews in addition to higher voted old dated reviews. The first step was to conduct a literature survey in order to get familiar with the work that had been done in the domain of text classification including the work relevant to the review helpfulness measurement. It was found that the work done regarding the review helpfulness focused mostly on finding a correlation between review

content-centric features and the helpfulness measure of the review. It was also noticed that the word embedding based features had been used in the domain of text classification other than that of review helpfulness.

## **2.3 Objectives**

The aim of this project is to investigate if sentimental analysis is feasible for the classification of product reviews from Amazon.com. Therefore, we will compare the performance of different classification algorithms on the binary classification (positive vs. negative) of product reviews from Amazon.com. Thereby, we want to investigate whether the category of products the reviews come from influence the performance of this classification. Once found the best performing classifier, it will be applied on new Amazon.com datasets containing reviews of different product categories and these results will be compared.

## Chapter 3

### SCRUM MEETINGS

#### **On 17-11-2021**

On this day I started searching the miniproject topic based on the new technology such as deep learning, IoT, machine learning, classification, prediction etc”.

#### **On 15-12-2021**

The topic was selected and did the detail study of the topic, the required dataset was selected. The dataset was searched from the different site such as kaggle, dataset etc.

#### **On 17-12-2021**

This day I submitted the synopsis and research paper to guide for the topic approval.

#### **On 20-12-2021**

After getting approval from the guide, the algorithm and model for the project were structured. Then the algorithm were chosen.

**On 21-12-2021**

On this day mam took a detailed class on how to do the project, what IDEs to use, what paper are referred, what steps are follow to do the project and so on

**On 22-12-2022**

According to the project the required IDE such as Visual Studio Code, Colab are choosen. Even checked whether the system was efficient to train the model. Here colab to code the project, then started to deploying the model using the algorithm. Python language is used to code the project.

**On 24-12-2022**

After the project first review according to Mam's opinion added two new Algorithm to the project to find which algorithm is having more accuracy rate.

**On 29-01-2022**

Used different algorithm/data model then choose the maximum accuracy one. The algorithm used are:-

Multinomial Naive Bayes

**On 17-01-2022**

Started to do project coding. Firstly study the dataset and download the dataset from kaggle. The dataset is about Reviews from Amazon websites

**On 19-01-2022**

Testing the data application

**On 24-01-2022**

The training done is the data model then choose the maximum accuracy with naive bayes for predicting the review. Multinomial Naive Bayes model is used for prediction.

**On 02-02-2022**

Created the git repository.

**On 07-02-2022**

Used flask for connection.

## Chapter 4

# IMPLEMENTATION

The aim of this project is to investigate if sentimental analysis is feasible for the classification of product reviews from Amazon.com. Therefore, we will compare the performance of different classification algorithms on the binary classification (positive vs. negative) of product reviews from Amazon.com. Thereby, we want to investigate whether the category of products the reviews come from influence the performance of this classification. Once found the best performing classifier, it will be applied on new Amazon.com datasets containing reviews of different product categories and these results will be compared. The model uses multinomial naive bayes algorithm for developing the model. And different featurization vector such as Bag of Words (BOW) and TFIDF vectors are applied on these models to get more accurate values.

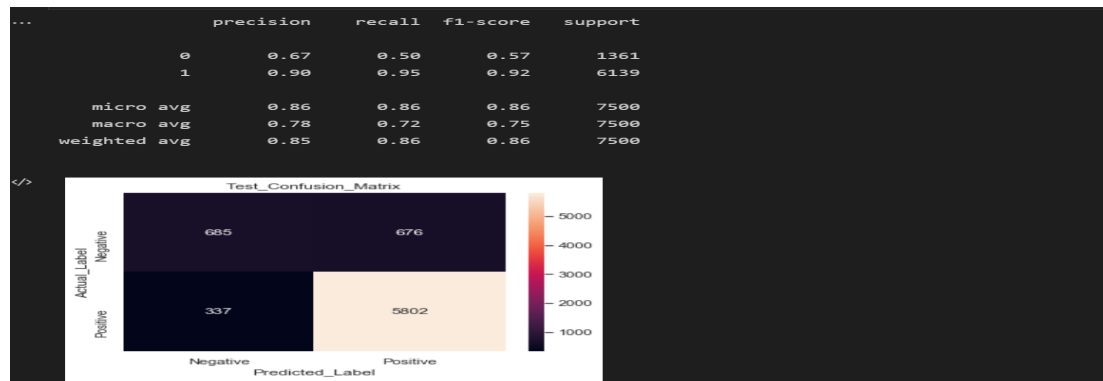
- Feature Selection: Finding out the best feature which will contribute and have good relation with target variable.
- naive bayes: Finding the important.
- Hyperparameter Tuning: Hyperparameter tuning is choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a model argument whose value is set before the learning process begins.

## ALGORITHM

### Naive Bayes

The naive Bayes Algorithm is one of the popular classification machine learning algorithms that helps to classify the data based upon the conditional probability values computation. It implements the Bayes theorem for the computation and used class levels represented as feature values or vectors of predictors for classification. Naive Bayes Algorithm is a fast algorithm for classification problems. This algorithm is a good fit for real-time prediction, multi-class prediction, recommendation system, text classification, and sentiment analysis use cases. Naive Bayes Algorithm can be built using Gaussian, Multinomial and Bernoulli distribution. This algorithm is scalable and easy to implement for a large data set.

The Multinomial Naive Bayes algorithm is a Bayesian learning approach popular in Natural Language Processing (NLP). The program guesses the tag of a text, such as an email or a newspaper story, using the Bayes theorem. It calculates each tag's likelihood for a given sample and outputs the tag with the greatest chance.



## 4.1 System Architecture

The use case diagram that describes the operation of the system .

## 4.2 Dataset

The data requirements is very high for the project. We get a dataset that contains the component like:-

- Id
- ProductId - unique identifier for the product
- UserId - unique identifier for the user
- ProfileName
- HelpfulnessNumerator - number of users who found the review helpful
- HelpfulnessDenominator - number of users who indicated whether they found the review helpful
- Score - rating between 1 and 5
- Time - timestamp for the review
- Summary - brief summary of the review
- Text - text of the review

## 4.3 Modules

### 4.3.1 Data Preprocessing

In Machine Learning, Data pre-processing is one of the most integral steps before applying algorithms. Because Machine Learning algorithms do not work with raw data, so text data needs to be cleaned and converted into numerical vectors. This process is called text-processing. These are below basic steps that we are going to show you in this paper. 1) Understanding the data: First of all, you need to see what the data is all about and what parameters (Stopwords, Punctuation, html

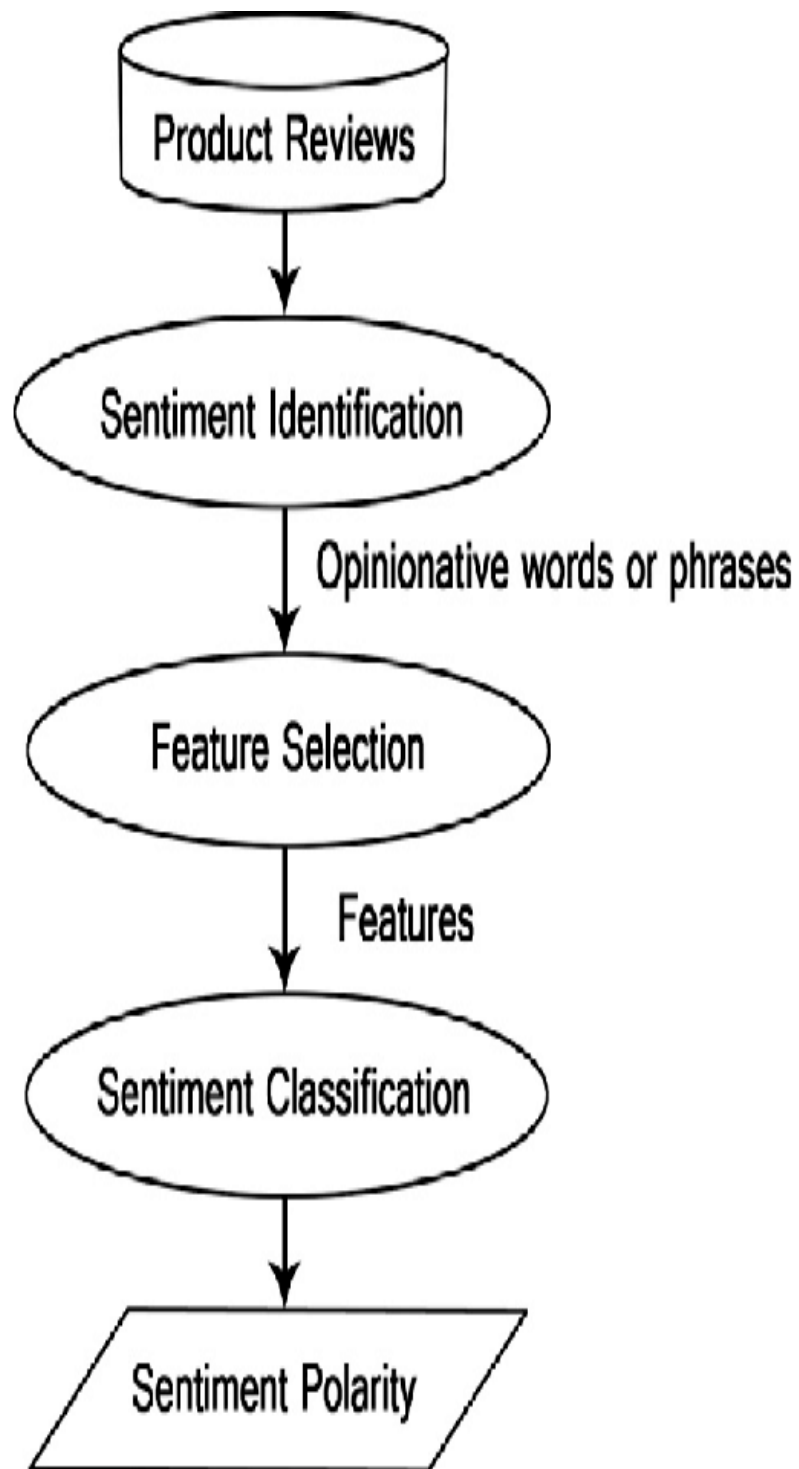


tag.... etc) are in the data. 2) Data Cleaning: In this step, we are going to discuss about how to remove parameters. 3) Techniques for encoding text data: There are lots of techniques for encoding text data. But below are the techniques I have mostly used while solving real-world problems.

### **4.3.2 TEXT ENGINEERING**

Feature Selection: Finding out the best feature which will contribute and have good realtion with target variable.

1. multinomial nb: Finding the important.
2. Hyperparameter Tuning:Hyperparameter tuning is choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a model argument whose value is set before the learning process begins.



## **Chapter 5**

### **RESULT ANALYSIS**

The Review Prediction helps us to know about the product. Whether we should buy the product or not also helps the manufactures to improve the product if not good. The output will be either positive or negative

## **Chapter 6**

# **CONCLUSION AND FUTURE SCOPE**

### **6.1 Conclusion**

In this study, the focus was on building an automated text classification system which can predict the helpfulness measure of an online review irrespective of the time of posting. The purpose of this was to provide both consumers and manufacturers a wide variety of reviews to choose from by including the most recent yet unvoted reviews in addition to higher voted old dated reviews. The first step was to conduct a literature survey in order to get familiar with the work that had been done in the domain of text classification including the work relevant to the review helpfulness measurement. It was found that the work done regarding the review helpfulness focused mostly on finding a correlation between review content-centric features and the helpfulness measure of the review. It was also noticed that the word embedding based features had been used in the domain of text classification other than that of review helpfulness. I.

## **6.2 Future Scope**

In the future, if more data could be accessed such as the current availability of seats, the predicted results will be more accurate.

## **Chapter 7**

### **SOURCE CODE**

```
#define a function for splitting of data
from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.3,shuffle=False)
print(X_train.shape,Y_train.shape)
print(X_test.shape,Y_test.shape)
```

(17500,) (17500,)

(7500,) (7500,)

Python

Figure 7.1: Train Data

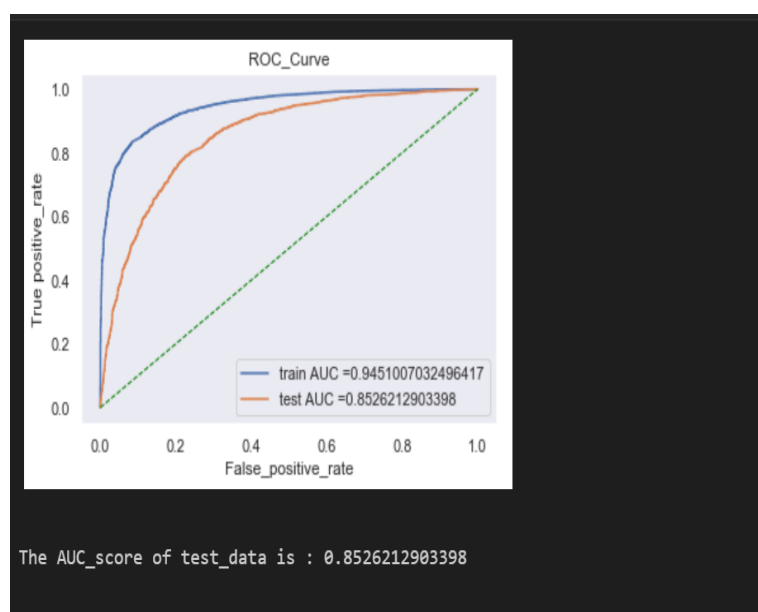


Figure 7.2: ROC Curve

```
#define a function for splitting of data
from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.3,shuffle=False)
print(X_train.shape,Y_train.shape)
print(X_test.shape,Y_test.shape)
```

Python

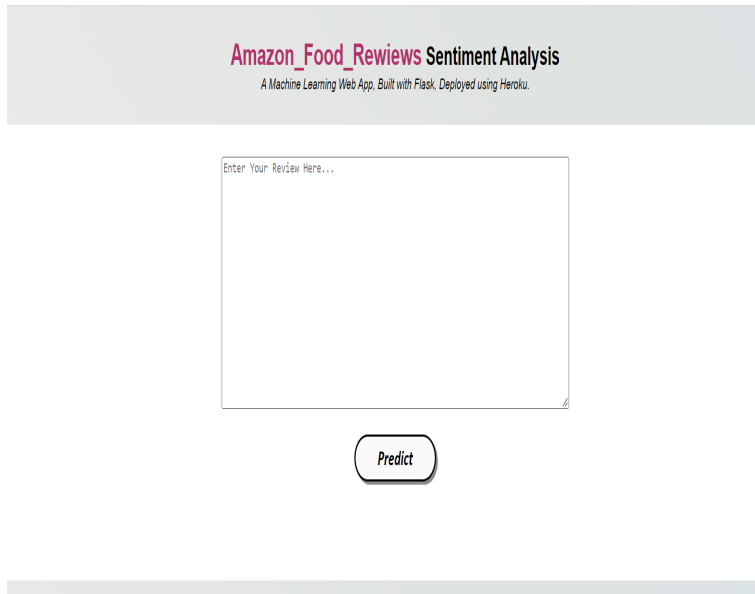
```
(17500,) (17500,)
(7500,) (7500,)
```

Figure 7.3: Splitting and Fitting data



## **Chapter 8**

### **SCREEN SHOTS**



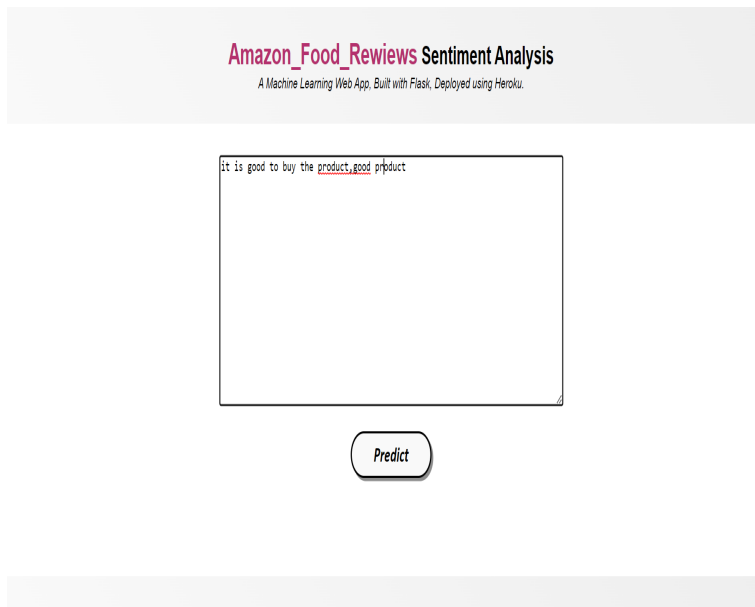
Amazon\_Food\_Rewiews Sentiment Analysis

A Machine Learning Web App. Built with Flask, Deployed using Heroku.

Enter Your Review Here...

Predict

Figure 8.1: Main Page



Amazon\_Food\_Rewiews Sentiment Analysis

A Machine Learning Web App. Built with Flask, Deployed using Heroku.

It is good to buy the product, good product

Predict

Figure 8.2: Prediction Page



Figure 8.3: Prediction Page

## Chapter 9

## REFERENCES

- [www.youtube.com](http://www.youtube.com)
- [www.wikipedia.com](http://www.wikipedia.com)
- 1 Deepu S, Pethuru Raj, and S.Rajaraajeswari. “A Framework for Text Analytics using the Bag of Words (BoW) Model for Prediction” International Journal of Advanced Networking Applications (IJANA).
- [2]Kim, S. M., Hovy, E. (2004, August). Determining the sentiment of opinions. In Proceedings of the international conference on Computational Linguistic
- [3] <https://www.kaggle.com/nikhilmittal/Amazon-Fine-Food-Review>