



Tribhuvan University

Faculties of Humanities and Social Sciences

**SENTIMENT ANALYSIS OF PRODUCT REVIEWS USING
NAÏVE BAYES
A PROJECT REPORT**

Submitted to

Department of Computer Application

Deerwalk Institute of Technology, Kathmandu

*In partial fulfillment of the requirements for the Bachelor in Computer
Application*

Submitted By

Kristina Maharjan

August, 2022

Under the supervisor of

Hitesh Karki



Tribhuvan University

Faculties of Humanities and Social Sciences

Deerwalk Institute of Technology

SUPERVISOR'S RECOMENDATION

I hereby recommend that this project prepared under my supervision by KRISTINA MAHARJAN entitled “**SENTIMENT ANALYSIS OF PRODUCT REVIEWS USING NAÏVE BAYES**” in partial fulfillment of the requirements for the bachelor degree in Computer Application be processed for the evaluation.

.....

SIGNATURE

Hitesh Karki

SUPERVISOR

Campus Chief

Department of Computer Application

Deerwalk Institute of Technology

Sifal, Kathmandu



Tribhuvan University

Faculties of Humanities and Social Sciences

Deerwalk Institute of Technology

LETTER OF APPROVAL

This is to certify that this project prepared by KRISTINA MAHARJAN entitled **“SENTIMENT ANALYSIS OF PRODUCT REVIEWS USING NAÏVE BAYES”** in partial fulfillment of the requirements for the bachelor degree in Computer Application has been well studied. In our opinion it is satisfactory in the scope and quality as a project for the required degree.

| | |
|--|--|
| <p>.....</p> <p>Mr. Hitesh Karki Supervisor DWIT College</p> | <p>.....</p> <p>Mr. Hitesh Karki Chairman DWIT College</p> |
| <p>.....</p> <p>Mr. Ritu Raj Lamsal Project Coordinator DWIT COLLEGE</p> | <p>.....</p> <p>Mr. Bhoj Raj Joshi External Examiner Senior Lecturer Patan Multiple Campus</p> |

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to several individuals for supporting me throughout my project. First, I would like to express my appreciation to my supervisor, Hitesh Karki, for his enthusiasm, patience, insightful comments, helpful information, practical advice, and unceasing ideas that have helped me tremendously at all times in my project.

Also, I would like to appreciate my family members, who have always supported me and heartfelt thanks to my friends and seniors who have helped me with several ideas in this project.

Kristina Maharjan

Roll No:- 11751303

Date:- August, 2022

ABSTRACT

With the ever-growing technological advancements, internet is the most valuable source for collecting data, reviews for products and services. It is very difficult to extract and understand such reviews. Sentiment analysis helps to understand and extract opinions from the given review.

In this project, a system has been built for the prediction of sentiment of product reviews. In order to extract reviews, web crawler, Naïve Bayes have been used. Similarly, machine learning has been used for reviews classification purpose. An authentic site has been chosen in order to implement these classifications.

For training purpose, 1000 datasets have been used. Similarly, for the test purpose, 204 datasets have been used. The test showed the accuracy of 72% using Naïve Bayes. This paper provides an overall survey about sentiment analysis related to product reviews.

Keywords: *Sentiment analysis, Machine learning, Classification, Training, and Testing data*

TABLE OF CONTENTS

| | |
|--|------|
| SUPERVISOR’S RECOMMENDATION | ii |
| LETTER OF APPROVAL | iii |
| ACKNOWLEDGEMENT | iv |
| ABSTRACT..... | v |
| TABLE OF CONTENTS..... | vi |
| LIST OF FIGURES | viii |
| LIST OF TABLES | ix |
| LIST OF ABBREVIATIONS..... | x |
| CHAPTER 1: INTRODUCTION | 1 |
| 1.1 Introduction | 1 |
| 1.2 Problem Statement | 1 |
| 1.3 Objectives..... | 2 |
| 1.4 Scope and Limitation | 2 |
| 1.5 Development Methodology..... | 2 |
| 1.6 Report Organization | 4 |
| CHAPTER 2: BACKGROUND STUDY AND LITERATURE REVIEW | 5 |
| 2.1 Background Study | 5 |
| 2.2 Literature Review | 5 |
| 2.2.1 Supervised machine learning for sentiment analysis | 6 |
| 2.2.2 Machine Learning Approach..... | 6 |
| CHAPTER 3: SYSTEM ANALYSIS AND DESIGN | 7 |
| 3.1 System Analysis | 7 |
| 3.1.1 Requirement Analysis..... | 7 |
| 3.1.2 Feasibility Analysis..... | 8 |

| | |
|--|----|
| 3.1.2.1 Technical Feasibility..... | 8 |
| 3.1.2.2 Economic Feasibility | 8 |
| 3.1.2.3 Operational Feasibility | 9 |
| 3.1.2.4 Schedule Feasibility | 9 |
| 3.1.3 Process Modelling: DFD | 10 |
| 3.2 System Design | 11 |
| 3.2.1 Architectural Design | 11 |
| 3.2.2 Interface Design | 16 |
| 3.3 Algorithm..... | 17 |
| 3.3.1 Naïve Bayes Algorithm..... | 17 |
| 3.3.2 Porter Stemming Algorithm..... | 19 |
| 3.3.3 Term Frequency | 20 |
| CHAPTER 4: IMPLEMENTATION AND TESTING | 21 |
| 4.1 Implementation..... | 21 |
| 4.1.1 Tools and Technologies | 21 |
| 4.1.2 Implementation Details of Modules..... | 22 |
| 4.2 Testing..... | 22 |
| 4.2.1 Manual Testing | 22 |
| 4.2.2 Accuracy of the application | 23 |
| 4.2.2 Overall System Design | 24 |
| CHAPTER 5: CONCLUSION AND FUTURE RECOMMENDATION..... | 25 |
| 5.1 Conclusion..... | 25 |
| 5.2 Outcome | 25 |
| 5.3 Future Recommendations..... | 25 |
| REFERENCES | 26 |
| APPENDIX..... | 27 |

LIST OF FIGURES

| | |
|---|----|
| Figure 1: Waterfall Model..... | 3 |
| Figure 2: Use-case Diagram..... | 7 |
| Figure 3: Gantt Chart..... | 9 |
| Figure 4: Level 0 DFD..... | 10 |
| Figure 5: Level 1 DFD..... | 10 |
| Figure 6: Flow Diagram of Overall System..... | 11 |
| Figure 7: Flow Diagram of Crawler for training Dataset..... | 12 |
| Figure 8: Flow Diagram of Preprocessing..... | 13 |
| Figure 9: Flow Diagram of Feature Extraction..... | 14 |
| Figure 10: Flow Diagram of Crawler designed to Fetch Test Data..... | 15 |
| Figure 11: Homepage Interface Design..... | 16 |
| Figure 12: Result Page Interface Design..... | 17 |
| Figure 13: Implementation of Porter Stemming..... | 19 |
| Figure 14: Homepage..... | 27 |
| Figure 15: Sentiment Analysis of the Product..... | 28 |
| Figure 16: Product Link..... | 28 |
| Figure 17: Graphical Representation..... | 29 |
| Figure 18: Training Dataset..... | 29 |
| Figure 19: Test Dataset..... | 30 |

LIST OF TABLES

| | |
|---------------------------|----|
| Table 1: Test Case 2..... | 23 |
| Table 2: Test Case 3..... | 24 |

LIST OF ABBREVIATIONS

| | |
|------|---|
| HTML | Hypertext Markup Language |
| CSS | Cascading Style Sheet |
| NM | NumPy |
| PD | Pandas |
| TFID | Term Frequency-Inverse Document Frequency |

CHAPTER 1: INTRODUCTION

1.1 Introduction

Sentiment analysis is a topic of great interest and development due to higher uses of social media opinions, messages and reviews are generated and shared in uncontrollable amount; analysis of these data can make greater difference in product delivery, customer satisfaction and business survival.

Sentiment analysis of product reviews has recently become very popular in text mining and computational linguistics research. Sentiment analysis at present context has been a hyped topic. There are various algorithms that can be applied to achieve the task. However, the nature of the data is a major factor to be considered while choosing an algorithm.

The application is about sentiment analysis of individual product reviews. Taking into consideration the data source and type, Naïve Bayes algorithm has been implemented in the application for sentiment analysis. This project will be used by vendors to extract invaluable data from reviews they receive in their product listings and aiding them with market analysis, brand monitoring, and more.

1.2 Problem Statement

Everyday millions of data is being collected on e-commerce sites which contains people opinion about many things like manufacturing products, quality, quantity, etc. These texts are usually difficult and time-consuming to analyze, understand, and sort through. Many companies want to know how positive or negative peoples are about their product.

There are multiple products of vendors in multiple stores which makes the vendor difficult and tedious to extract and check every single review. For example, a Nepali brand, KTMCTY has their products in multiple stores like Daraz, SastoDeal. It will be hard for them to extract the reviews of every product. Sentiment Analysis of Product

Reviews Using Naïve Bayes Algorithm will provide the positive or negative sentiment on review of products.

1.3 Objective of the Project

The objective of this project are:

- To enable the vendors to identify the reviews and distinguish good reviews from bad ones.
- To analyze and categorize reviews.
- To help vendors monitor brand and product sentiment in customer feedback, and understand customer needs.
- To help vendors to know the expectations before launch of product.

1.4 Scope and Limitation

Product review sentiment analysis can be used by the vendors to identify the reviews of the product. The vendors can differentiate the positive and negative reviews. The model is only valuable to the specific company whose data attributes matches with the data attributes that was used to build the model.

In this system, the multiple products cannot be crawled at the same time, only a single product is crawled at a time, single product reviews are extracted at a time. The products of Daraz is only useful. This application will not give 100% accurate result. The neutral value has not been classified. This system can only analyze the sentiment of English words.

1.5 Development Methodology

Waterfall model was brought into practice because the phase of model is processed and completed one at a time and most importantly, they do not overlap with each other.

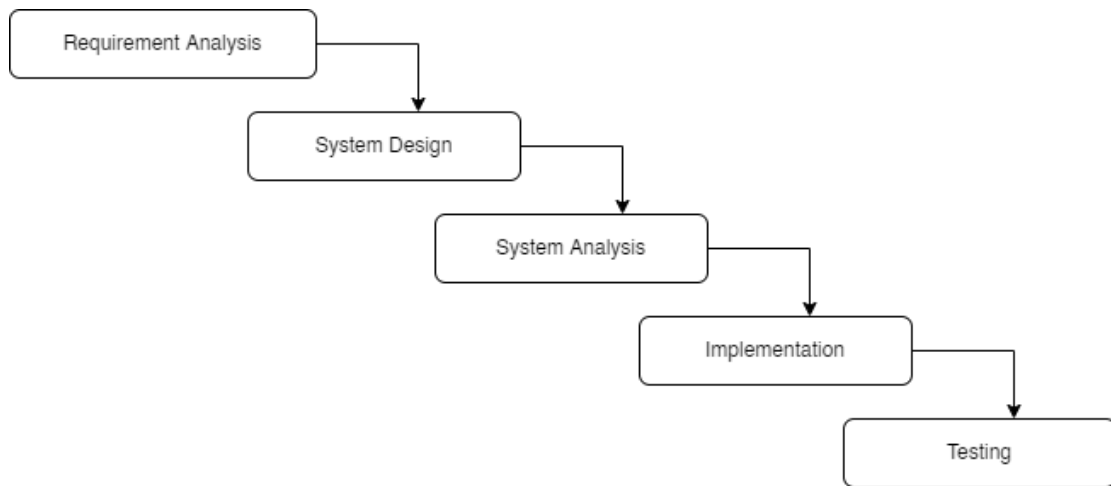


Figure 1 : Waterfall Model

Requirement Analysis

During this phase, detailed requirements of the system to be developed are gathered by doing research.

System Design

Then, outline of the project was made including the features that are required. The process modeling, architecture design, interface design, etc. were designed.

Implementation

The frontend has been developed using HTML, CSS and JavaScript. Python was used for the logical part which has made an extensive use of FLASK framework.

Testing

After the implementation was done, testing was performed to test the functionality of the application. Various test cases were carried out to check whether the feature works properly or not.

1.6 Report Organization

The report is organized into six chapters.

Chapter 1: Includes description about sentiment analysis of product reviews, problem statements, objectives, scopes and limitations.

Chapter 2: Consists of literature review about previous work done in related field.

Chapter 3: Comprises of requirement analysis. The requirement analysis further consists of functional and non-functional requirements. This section talks about the research done about sentiment analysis of product reviews, and the functional and non-functional requirements of this project. This chapter includes the algorithm that is used for this project.

Chapter 4: Consists of system design which includes system architecture, interface design, use-case diagram. This section includes diagrams that help to elaborate on the overall 3 design of the system proposed in this project.

Chapter 5: Includes conclusion. This section includes a conclusion to this paper.

Chapter 6: Includes limitation and future enhancements.

CHAPTER 2: BACKGROUND STUDY AND LITERATURE REVIEW

2.1 Background Study

With the rapid growth of online shopping platforms, more and more customers intend to share their shopping experience and product reviews on the Internet.

Sentiment is positive or negative reviews about product or on any topics. We people can identify reviews by reading whether it is positive or negative. But if there is huge data to be read then it would be tedious and time consuming. So, if all this process could be done with the help of automated program, then it would be easier and above manual process could be eliminated.

Firstly, stop words are removed from each news article. Predefined list of stop words in English is used as reference [6]. Then, porter stemming algorithm is applied on the article in order to represent words with similar meaning as a single word. Porter stemmer algorithm is the process for removing the common morphological and in flexional endings from words in English. Finally, classification algorithm generates either the reviews are positive or negative.

As manually entering the reviews take a lot of effort, a search button by clicking on which it takes us to a product website URL from which the positive and negative reviews is determined and shown in the screen. We have developed a system where reviews extracted will be automatically classified as positive or negative using Naive Bayes algorithm.

2.2 Literature Review

The purpose of this project study is to classify the sentiment of people through opinion. In this chapter, the major ground work and preliminaries related to the subject of the study, is review. The various related approaches and review of our project is review in this chapter.

2.2.1 Supervised machine learning for sentiment analysis

Sentiment analysis have become the growing area in the natural language processing. Supervised machine learning algorithm like Naïve Bayes algorithm paly vital role in the sentiment analysis. There are many researched carried out for sentiment analysis.

(Waykar, Wadhwani, Pooja, & Kollu, 2016) Have focused mainly on the Naïve Bayes classifier. They take the baseline for their research as (Pang, Lee, & Vaithyanathan, 2002). They display the result on pie chart for positive and negative for the specific keyword.

Classification of text on the basis of sentiment and genre initiated form the early works and research of Hearst and Kessler (1992). The main aim of the current classifications systems is to identify the opinion of words, phrases, sentences or a document as a whole so as to classify them as positive [1] [3], negative or neutral. In advanced form sentiment analysis can be used to identify other emotions like happy, sad, anger, fear, and disgust among other which requires more in-depth analysis of the content.

2.2.2 Machine Learning Approach

[3]Machine learning approach for sentiment analysis has the capability to model multiple features than the symbolic approach. This makes the algorithm easier to adapt to changing input and also makes it possible to measure the degree of uncertainty by which a classification is made. The supervised methods that uses example manually classified by the humans are more in used and are the most common among the machine learning methods.

CHAPTER 3: SYSTEM ANALYSIS AND DESIGN

3.1 System Analysis

Requirement Analysis was done for the application by studying and researching the related projects available on the web and following requirements were collected.

3.1.1 Requirement Analysis

i) Functional Requirement

The functional requirements are set out as follows:

- The application shall be able to classify positive and negative reviews.
- The application shall be able to determine which products are customer centric.

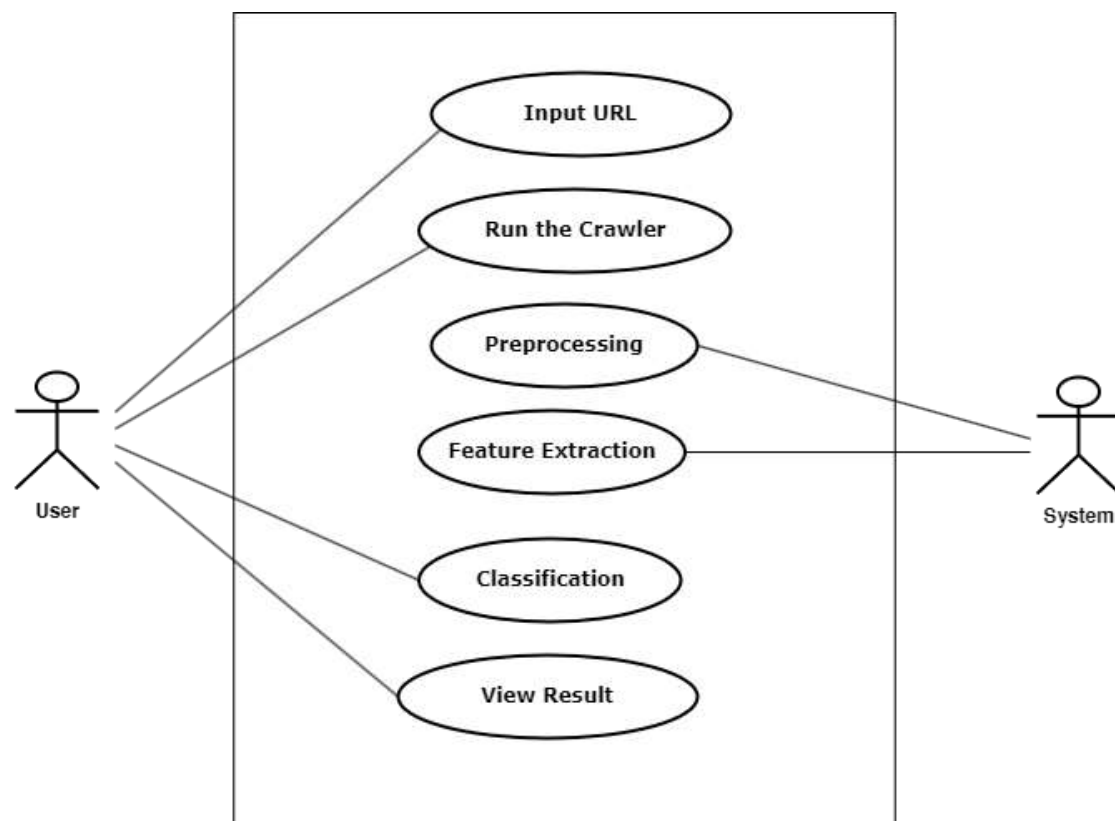


Figure 2 : Use-case diagram

ii) Non-Functional Requirement

The non-functional requirements are set out as follows:

- The application must respect the customer's privacy.
- Websites from which data is to be extracted must be crawler friendly.
- The internet speed should be fast.
- Operation of the application must not be time and resource consuming.

3.1.2 Feasibility Analysis

The main objective of the feasibility study is to test the technical, economic, operational, and schedule feasibility of the system.

3.1.2.1 Technical Feasibility

This project validates the technical resources and capabilities to convert the ideas into working system. It was built with the help of HTML, CSS, and JavaScript for front-end and Python for back-end engine development. The algorithm used for this project is Naïve Bayes Algorithm. These technologies were chosen because they could deliver the outcome, I expect my project to deliver. The application can be declared technically feasible as all the technical resources are easily available and accessible.

3.1.2.2 Economic Feasibility

Economic feasibility involves a cost benefits analysis to identify how well, or how poorly this project will be concluded. Proposed system requires development tools and software such as PyCharm which are free of cost and is easily available on internet. The project is economically feasible.

3.1.2.3 Operational Feasibility

Operational feasibility is to be taken as an integral part of the project implementation. Proposed projects are beneficial only if they can meet the user's operating requirements. It is based on client-server architecture and needs the internet connection to access information. Hence, it is understood that sentiment analysis of product reviews is an operationally feasible application.

3.1.2.4 Schedule Feasibility

As we can see in the Gantt chart, the system development is completed within the required time frame.

| Activities | Week 1 | Week 2 | Week 3 | Week 4 | Week 5 | Week 6 | Week 7 |
|-------------------------------------|--------|--------|---------|---------|---------|---------|---------|
| Planning, Research | | | | | | | |
| Crawl data using Selenium | | | | | | | |
| Train data | | | | | | | |
| Test Data | | | | | | | |
| Documentation | | | | | | | |
| | | | | | | | |
| Activities | Week 8 | Week 9 | Week 10 | Week 11 | Week 12 | Week 13 | Week 14 |
| Test Data | | | | | | | |
| Naive Bayes Implementation | | | | | | | |
| Front end Implementation with flask | | | | | | | |
| Documentation | | | | | | | |
| Review And Final Report | | | | | | | |

Figure 3 : Gantt Chart

3.1.3 Process Modelling: DFD



Figure 4 : Level 0 DFD

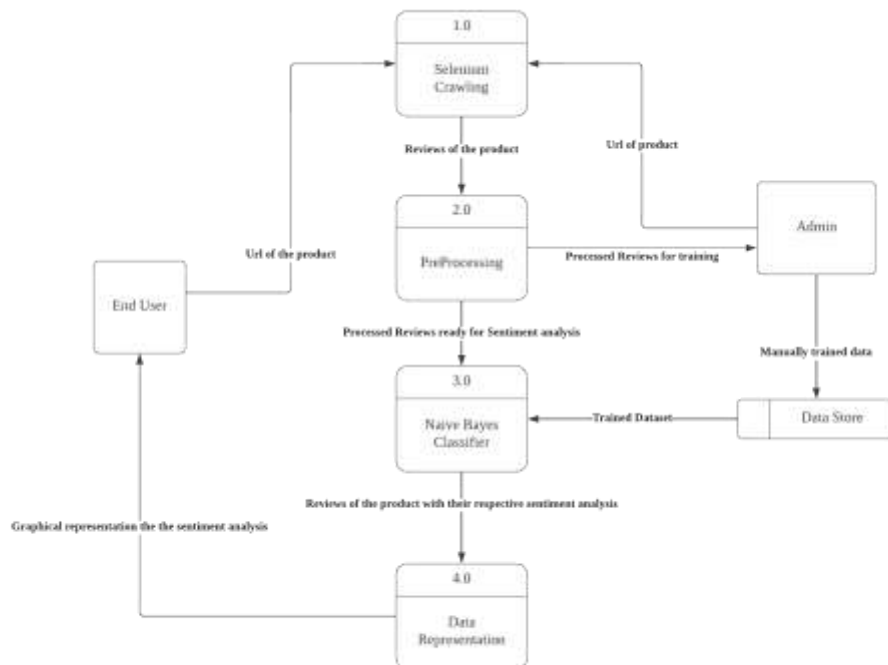


Figure 5 : Level 1 DFD

3.2 System Design

System design for the project includes the logical design of the application. System design is the step after analyzing the system to provide a pictorial view of the system being developed. It is the process of defining the elements of a system such as the architecture, modules and components, the different interfaces of those components and the data that goes through that system.

3.2.1 Architectural Design

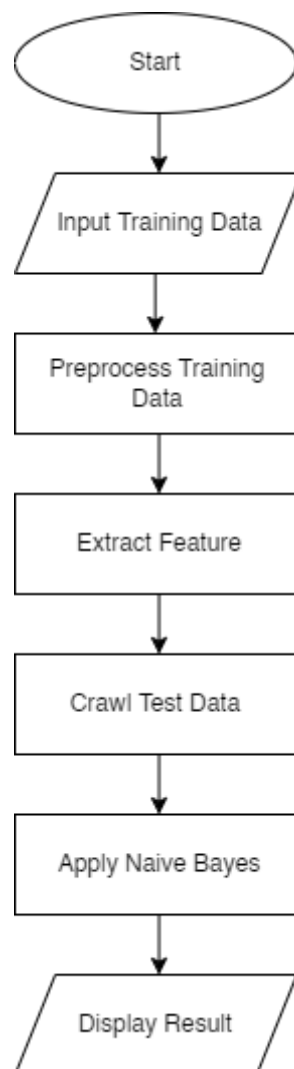


Figure 6 : Flow Diagram of Overall System

The above figure shows the abstract flow diagram of the system design for the project. The application begins by feeding the trained data to the system. This data is then preprocessed. Pre-processing step includes removing punctuations, stop words and non-alphabets. The next step is feature extraction. TFID and Count Vectorizer are implemented in this step. Then we continue by crawling the test data. These data comprise the headlines that are published on the current date.

Naïve Bayes is then applied to fit the data in the curve which gives the required result.

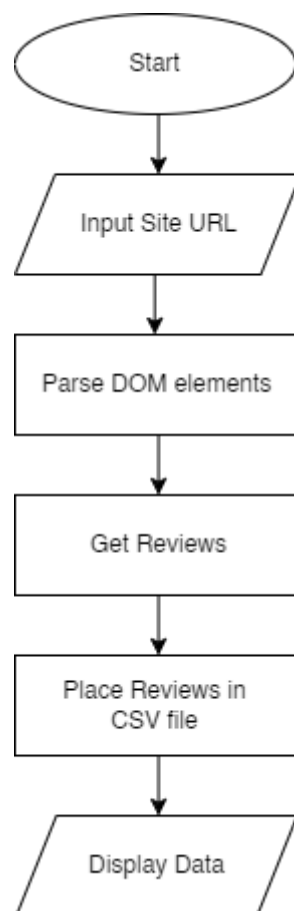


Figure 7 : Flow Diagram of Crawler for Training Dataset

As shown in above figure, for the crawler to implement, first the URL of input site must be fed into the system. By parsing DOM elements, the crawler goes into the

part of the site that is required for data extraction. The headlines are then fetched. The headlines are then extracted as csv file for display and weight assign.

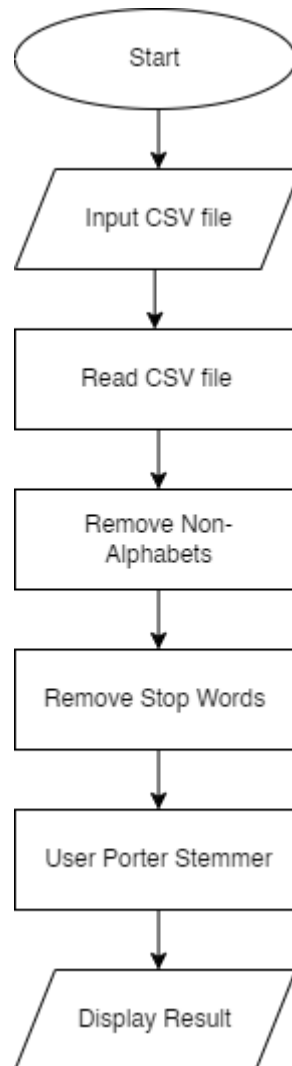


Figure 8 : Flow Chart of Preprocessing

The first step of preprocessing is to get the input file, as shown in above figure. The file is then opened to remove non-alphabets and stop words. Then the Porter Stemming Algorithm is applied to display a data set that is to be used for analysis.

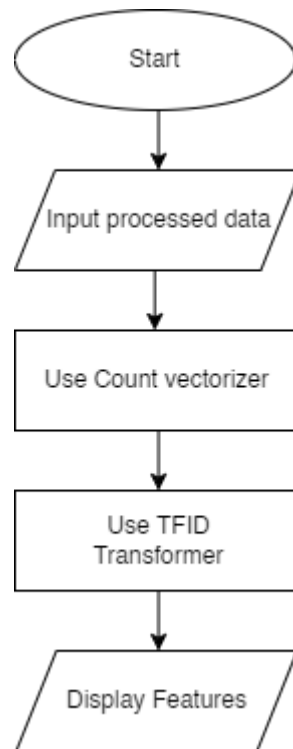


Figure 9 : Flow Chart of Feature Extraction

The next step is feature extraction, shown in the above. Feature extraction begins by processing the input and implementing the count vectorizer function. TFID is then used to display the features.

In order to use the test data, we need another crawler, shown in the below figure. This crawler crawls data, which is the headlines of news and then checks if it has been published on the current date. If the headline is not new, the process must end. However, if the headline has today's date, it fetches the headline and the analysis is performed.

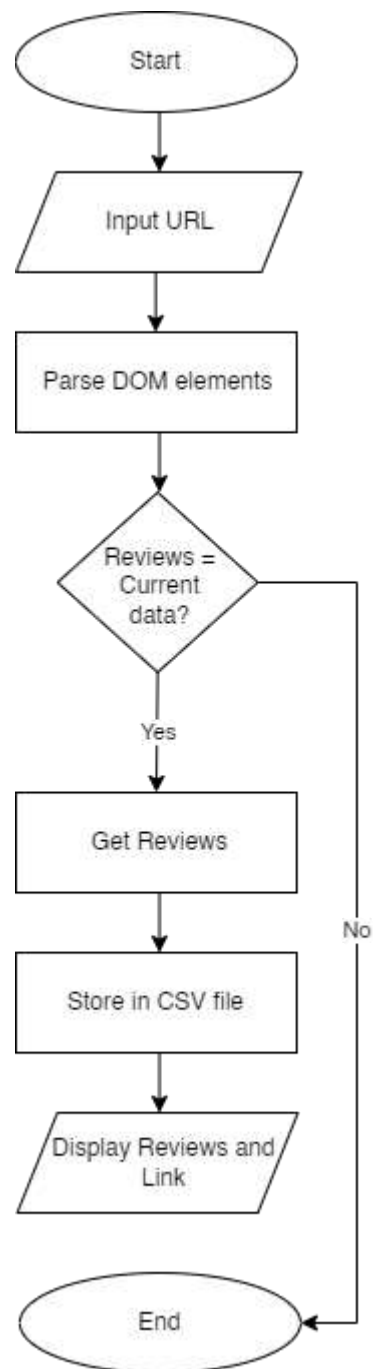


Figure 10 : Flow Chart of Crawler Designed to Fetch Test Data

3.2.2 Interface Design

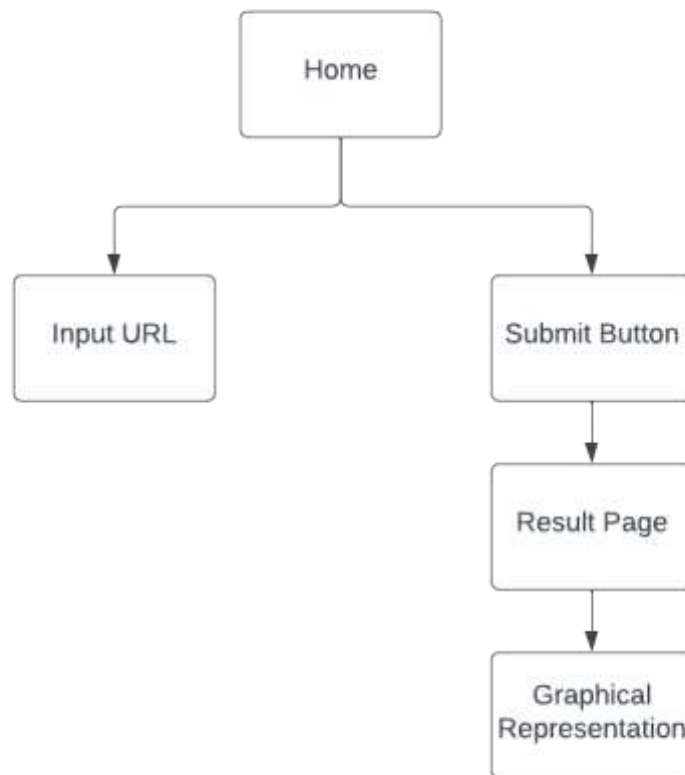


Figure 11 : Homepage interface design

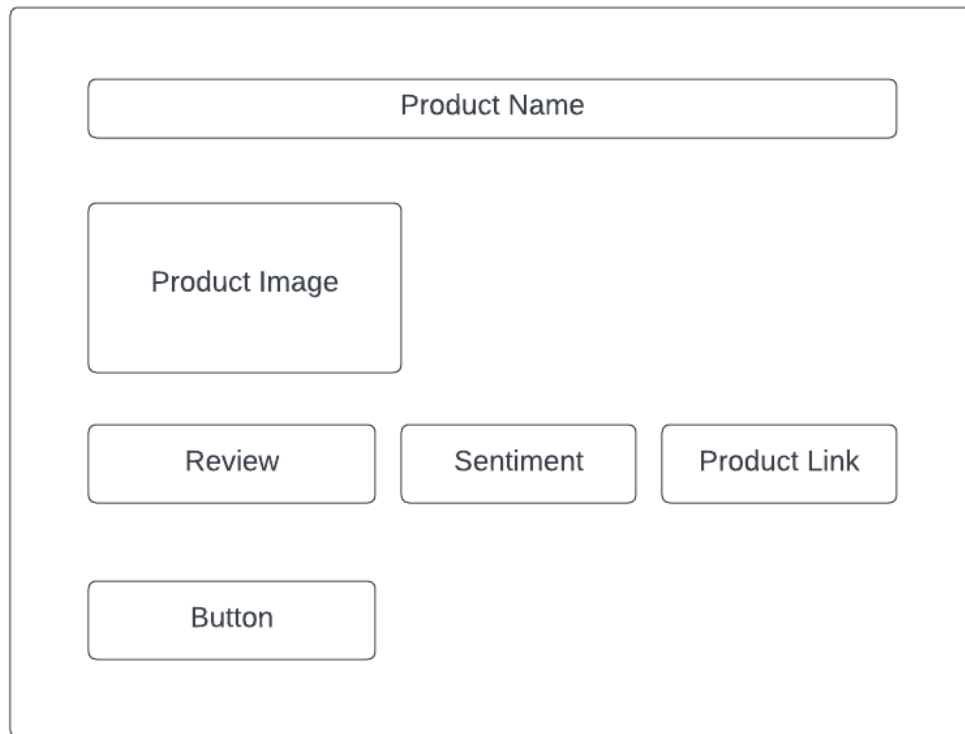


Figure 12 : Result page interface design

3.3 Algorithm

3.3.1 Naïve Bayes Algorithm

Naïve Bayes Theorem implements a classifier which assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. The Naïve Bayes model has been implemented in the project as it is easy to build and particularly useful for very large data sets. Along with simplicity, Naïve Bayes is known to outperform even highly sophisticated classification methods (Reference). In other words, it is a simple technique for constructing classifiers that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle. All of the principles are Naïve

Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

In the project, the algorithm has been implemented as follows:

First of all, the data set comprising reviews of product was converted into a frequency table. We have a csv file containing a product review. Each row in this dataset contains the review and whether the content is positive or negative. The positive news is given value 1 while the negative is given 0.

In order to do this, we'll train the algorithm using the reviews and classifications in train.csv, and then make predictions on the reviews in test.csv. The next step would then be to calculate the error using the actual classifications in test.csv, and evaluate how good the predictions were. Then, the likelihood table was generated using probability.

$$p(C_k | \mathbf{x}) = \frac{p(C_k) p(\mathbf{x} | C_k)}{p(\mathbf{x})}$$

Where,

- K is each possible outcome or classes. 2 in our case.
- P(C) is probability of each category being true. It is known as prior probability.
- P(X) is the probability of the event (occurrence of tokens).
- P(X/C) is the probability of the evidence (X) given that the hypothesis(C) is true.

P(C/X) is the probability of the hypothesis (C) given that the evidence(X) is there.

3.3.2 Porter Stemming Algorithm

Porter Stemming Algorithm is a process adapted to remove the commoner morphological and inflectional endings from words in English.

- The algorithm begins by first removing the plurals (*girls*, *puppies*) and past (*happened*) in the words.
- If the stem has vowel in it and the letter that precedes y is a consonant, the y in the end is turned into I (example: *happy*->*happi*)
- Double suffix such as *ization* is turned into a single suffix.
- Suffixes such as –full, -ness, -ant and –ance are dealt here
- First consonant sequence ‘m’ in the stem is looked at. If $m > 1$, final –e is removed and –ll is changed to –l.
- In the project, two preprocessing functions are built to remove stop words and non-alphabets.
- Then the function is passed to the Porter stemming Algorithm.

```
def process_review(review):
    # Remove hyperlinks
    review = re.sub(r'https?://\S+[\r\n]*', '', review)

    # Remove hashtags
    # Only removing the hash # sign from the word
    review = re.sub(r'#', '', review)

    # tokenize reviews
    review_tokens = word_tokenize(review)

    # Import the english stop words list from NLTK
    stopwords_english = stopwords.words('english')

    # Creating a list of words without stopwords
    clean_review = []
    for word in review_tokens:
        if word not in stopwords_english and word not in string.punctuation:
            clean_review.append(word)

    # Instantiate stemming class
    stemmer = PorterStemmer()

    # Creating a list of stems of words in review
    reviews_stem = []
    for word in clean_review:
        stem_word = stemmer.stem(word)
        reviews_stem.append(stem_word)
    return reviews_stem
```

Figure 13 : Implementation of Porter Stemming Algorithm

3.3.3 Term Frequency

Tf-idf stands for term frequency-inverse document frequency, and the tf-idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure that has been used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the tf-idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. One of the simplest ranking functions is computed by summing the tf-idf for each query term; many more sophisticated ranking functions are variants of this simple model.

Tf-idf can be successfully used for stop-words filtering in various subject fields including text summarization and classification. Typically, the tf-idf weight is composed by two terms: the first computes the normalized Term Frequency (TF) and the second term is the Inverse Document Frequency (IDF).

TF: Term Frequency is the number of times a word appears in a document, divided by the total number of words in that document which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length as a way of normalization.

$$TF(t) = (\text{Number of times term } t \text{ appears in a document}) / (\text{Total number of terms in the document}).$$

IDF: Inverse Document Frequency can be computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears. It measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and that may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms and scale up the rare ones.

$$IDF(t) = \log_e(\text{Total number of documents} / \text{Number of documents with term } t \text{ in it}).$$

Mathematically, $Tf-idf(t,d) = tf(t,d) \times idf(t,d)$

Here, the $tf(t,d)$ is the term frequency and $idf(t,d)$ is the inverse documents frequency.

CHAPTER 4: IMPLEMENTATION AND TESTING

4.1 Implementation

The project implies various development technologies for system development, HTML, CSS, JavaScript and Python are used for interface design and backend processing.

4.1.1 Tools and Technologies

The following technologies installed to setup the system used to build sentiment analysis:

CASE tools:

- a) Draw.io

Client side:

- a) HTML used to display the content in the browser.
- b) CSS is used to adjust the layout, look and design of the HTML content.
- c) Flask web framework is used for dynamic webpage generation and to display the predicted result in the browser as well as to handle page requests.
- d) JavaScript is used to program the behavior of web pages.

Server side:

- a) Python programming language is used to implement the core program logic.
- b) NumPy is to manipulate the large multidimensional arrays and matrices.
- c) Pandas is used for data manipulation and analysis.
- d) Naïve Bayes as machine learning algorithm

Sentiment Analysis of News has been built over Python for the logical part which has made an extensive use of FLASK framework. The frontend has been developed using HTML, CSS and JavaScript. These languages are used as it is easy to implement and is supported by every browser. HTML is used for presentation technology while CSS and JavaScript have been used to make the webpages more attractive and dynamic.

4.1.2 Implementation Details of Modules

The modules include

i) User Module

The user module includes aspects such as an interface for entering the url of product in order to obtain the sentiment of its reviews and an interface for viewing the results.

ii) System Module

System module includes three key components:

a) Crawl

Selenium crawls all the url and creates a dataset of its reviews.

b) Preprocessing:

Various data cleaning steps are performed on the given dataset of reviews.

c) Sentiment Analysis

The reviews are then fed into the system and classify the sentiment of each reviews as positive or negative using Naïve Bayes Algorithm.

4.2 Testing

Testing was conducted as it is an effective measure that should be carried out for efficient performance of the system. The test cases discussed below guarantees the quality of the product and helps in preventing from expensive costs as it safeguards the application from any possible failure.

During testing process, the headlines extracted from the crawler were used. The following four test cases were implemented to conduct testing.

Test Case – 1: Manual Testing

Test Case 2: Accuracy of the application

Test Case 3: Overall System Design

4.2.1 Manual Testing

In this testing the reviews data was first manually given a sentiment by reading the titles which was done by a general user. The reviews were given the value 0 or 1, 0 for bad reviews and 1 for good reviews. Then the result from the program was compared with the manually entered sentiment which yielded 72% for Naïve Bayes.

4.2.2 Accuracy of the application

The data has been imported in the system.

Table 1 : Test Case 2

| | |
|------------------|---|
| Precondition | Test and Trained data are given to the system. |
| Assumption | Weight is correctly assigned to all the data. |
| Test Steps | <ul style="list-style-type: none">• Sentiment of the trained data and predicted sentiment is matched.• Accuracy is calculated by the mean of predicted and obtained sentiment. |
| Expected Result | The result must separate reviews as good or bad. |
| Generated Result | An accuracy of 72% was obtained. |

4.2.3 Overall System Design

For overall system testing, 204 reviews were fed in the system. The result obtained from the test is shown below:

Table 2 : Test Case 3

| Correct Analysis | Incorrect Analysis |
|------------------|--------------------|
| 159 | 45 |

CHAPTER 5: CONCLUSIONS AND FUTURE RECOMMENDATIONS

5.1. Conclusions

Sentiment Analysis of Product Reviews was implemented using Flask framework of Python. The application benefits in evaluating sentiment of a reviews by placing a label on each review, eventually letting a user know whether the review is good or bad. Slow internet access can hinder the speed of the application however, excessive delay was not encountered. The reviews dataset was obtained by manually assigning weight to 1000 reviews which resulted in 72% accuracy to the application. Hence, this application has been built to provide users a choice of which products they prefer by simply letting them know if the products has something good or bad to offer.

5.2. Outcome

The system is designed to classify the reviews of the product as positive or negative. It is designed using different technology and is used by the vendors. This helps vendors monitor brand and product sentiment in customer feedback, and understand customer needs. The reviews of any product of Daraz can be extracted and classified.

5.3. Future Recommendations

More than one site can be used for analysis. Data set can be increased for further increase in accuracy. The project can be enhanced by accepting the entire review for analysis. Multiple product reviews can be extracted at a time.

REFERENCES

- [1] B. R. B. Wiebe J, "A corpus study of evaluative and speculative language.", 2000.
- [2] [Online]. Available: (<http://scikitLearn.org/CountVectorizer.html>)," .
- [3] P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of review," 2002.
- [4] Waykar, P., Wadhwani, K., Pooja, M., & Kollu, A. (2016). Sentiment Analysis of Twitter tweets using supervised. Int. Journal of Engineering Research and Applications.
- [5] S. C. F. Cane W. K. Leung, "Sentiment Analysis of Product Reviews.".
- [6] L. Pang, "Seeing stars: Exploiting class relationships for sentiment categorization," 2005.

APPENDIX



Figure 14 : Homepage



Figure 15 : Provided link for sentiment analysis


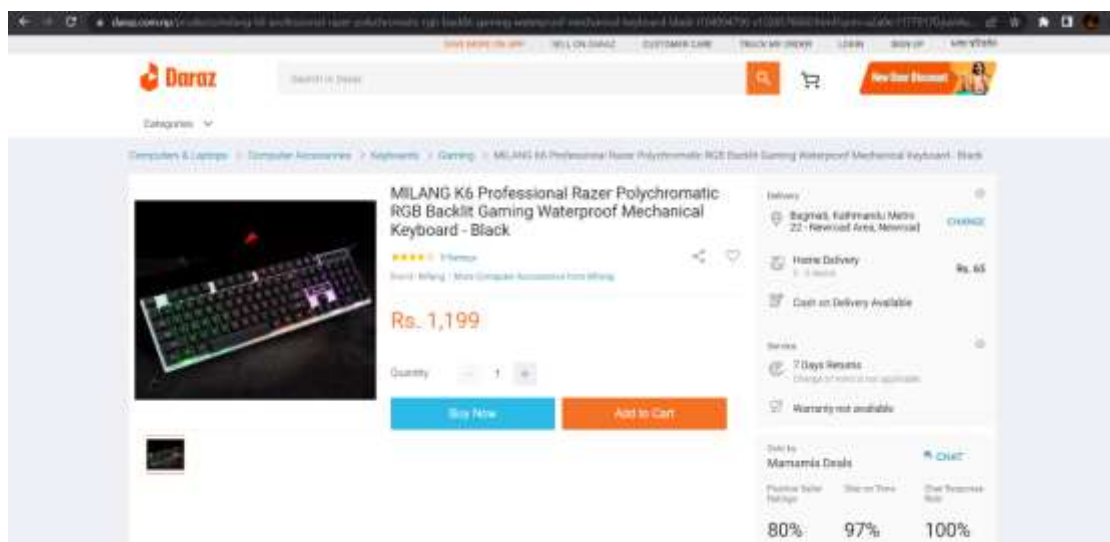
| MILANG K6 Professional Razer Polychromatic RGB Backlit Gaming Waterproof Mechanical Keyboard - Black Product Reviews | | |
|---|-----------|-------------------------|
|  | | |
| Review | Sentiment | Product Link |
| the keyboard light did not turn on automatically, is there a button or anything to turn it on? nothing mentioned on the box???? | Negative | Product |
| Nice keyboard its good for its price and took me a while to figure it out but we have to press scroll key on the keyboard to turn the lights on | Positive | Product |
| its cool it i ordered it the lights were functioning well for a couple of hours and then it stopped functioning please tell me what to do or i want an exchange or refund | Negative | Product |
| the product was as expected. Best for tight budget. | Positive | Product |
| Value for money products. I'm satisfied . | Positive | Product |
| it doesn't work Really disappointed | Negative | Product |
| nice | Positive | Product |
| Works fine | Positive | Product |
| View Graphical Representation | | |

Figure 16 : Sentiment analysis of the product



The screenshot shows the product page for the MILANG K6 Professional Razer Polychromatic RGB Backlit Gaming Waterproof Mechanical Keyboard - Black on the Daraz platform. The product is priced at Rs. 1,199. The page includes a search bar, navigation links, and a detailed product description. The delivery options section shows a delivery time of 2-3 days and a delivery fee of Rs. 65. The service section includes a 7-day return policy and a warranty that is not available. The page also features a 'Buy Now' button and an 'Add to Cart' button.

Figure 17 : Product Link

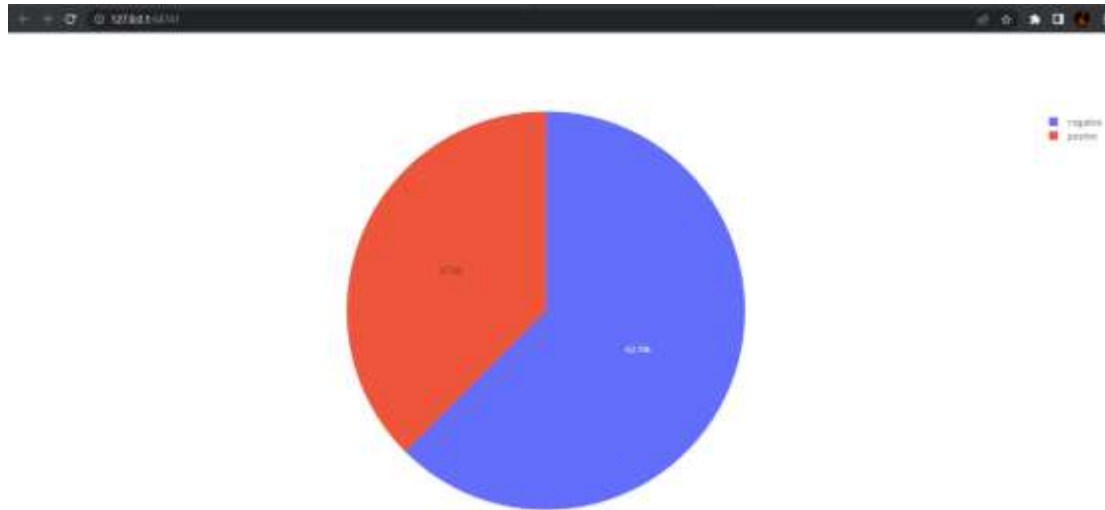


Figure 18 : Graphical Representation

```

url,review,rating,label
0,https://www.daraz.com.ng/products/4tm-cty-men-polyfiber-jacket-without-hoodie-kp105911-8a-1105022682-c1026620996.html?search=1,Wow, its just amazing",5
1,https://www.daraz.com.ng/products/4tm-cty-men-polyfiber-jacket-without-hoodie-kp105911-8a-1105022682-c1026620996.html?search=1,"delivered in just next 5
2,https://www.daraz.com.ng/products/4tm-cty-men-polyfiber-jacket-without-hoodie-kp105911-8a-1105022682-c1026620996.html?search=1,"delivered within 12 hour
3,https://www.daraz.com.ng/products/4tm-cty-men-polyfiber-jacket-without-hoodie-kp105911-8a-1105022682-c1026620996.html?search=1,Great quality ,3.1.0
4,https://www.daraz.com.ng/products/4tm-cty-men-polyfiber-jacket-without-hoodie-kp105911-8a-1105022682-c1026620996.html?search=1,Just received jacket... 5
5,https://www.daraz.com.ng/products/4tm-cty-men-polyfiber-jacket-without-hoodie-kp105911-8a-1105022682-c1026620996.html?search=1,Order large size but rece
6,https://www.daraz.com.ng/products/4tm-cty-men-polyfiber-jacket-without-hoodie-kp105911-8a-1105022682-c1026620996.html?search=1,quality is good fully sat
7,https://www.daraz.com.ng/products/4tm-cty-men-polyfiber-jacket-without-hoodie-kp105911-8a-1105022682-c1026620996.html?search=1,Its an great love it ,3.1
8,https://www.daraz.com.ng/products/4tm-cty-men-polyfiber-jacket-without-hoodie-kp105911-8a-1105022682-c1026620996.html?search=1,It was good...as experie
9,https://www.daraz.com.ng/products/4tm-cty-men-polyfiber-jacket-without-hoodie-kp105911-8a-1105022682-c1026620996.html?search=1,Satisfied with the qualiti
10,https://www.daraz.com.ng/products/4tm-cty-men-polyfiber-jacket-without-hoodie-kp105911-8a-1105022682-c1026620996.html?search=1,Good. According to price
11,https://www.daraz.com.ng/products/4tm-cty-men-polyfiber-jacket-without-hoodie-kp105911-8a-1105022682-c1026620996.html?search=1,satisfied so far.....3.
12,https://www.daraz.com.ng/products/4tm-cty-men-polyfiber-jacket-without-hoodie-kp105911-8a-1105022682-c1026620996.html?search=1,love this product.....3.
13,https://www.daraz.com.ng/products/4tm-cty-men-polyfiber-jacket-without-hoodie-kp105911-8a-1105022682-c1026620996.html?search=1,Good as expected,3.1.0
14,https://www.daraz.com.ng/products/4tm-cty-men-polyfiber-jacket-without-hoodie-kp105911-8a-1105022682-c1026620996.html?search=1,Also product ,3.1.0
15,https://www.daraz.com.ng/products/4tm-cty-men-polyfiber-jacket-without-hoodie-kp105911-8a-1105022682-c1026620996.html?search=1,Good service ,3.1.0
16,https://www.daraz.com.ng/products/4tm-cty-men-polyfiber-jacket-without-hoodie-kp105911-8a-1105022682-c1026620996.html?search=1,Great product ,3.1.0
17,https://www.daraz.com.ng/products/4tm-cty-men-polyfiber-jacket-without-hoodie-kp105911-8a-1105022682-c1026620996.html?search=1,good for me,3.
18,https://www.daraz.com.ng/products/4tm-cty-men-polyfiber-jacket-without-hoodie-kp105911-8a-1105022682-c1026620996.html?search=1,Nice,3.1.0
19,https://www.daraz.com.ng/products/4tm-cty-men-polyfiber-jacket-without-hoodie-kp105911-8a-1105022682-c1026620996.html?search=1,Thanks ,3.1.0
20,https://www.daraz.com.ng/products/4tm-cty-men-polyfiber-jacket-without-hoodie-kp105911-8a-1105022682-c1026620996.html?search=1,size,3.1.0
21,https://www.daraz.com.ng/products/8848-men-flaese-jacket-kp105757-8a-1103484168-c1024226058.html?search=1,"It was good he didnt about it. only reason
22,https://www.daraz.com.ng/products/8848-men-flaese-jacket-kp105757-8a-1103484168-c1024226058.html?search=1,"Color of the jacket was a little dark than t
23,https://www.daraz.com.ng/products/8848-men-flaese-jacket-kp105757-8a-1103484168-c1024226058.html?search=2,Product is good as expected. It's thin but it
24,https://www.daraz.com.ng/products/8848-men-flaese-jacket-kp105757-8a-1103484168-c1024226058.html?search=3,this is a great product.. well made... very co
25,https://www.daraz.com.ng/products/8848-men-flaese-jacket-kp105757-8a-1103484168-c1024226058.html?search=1,This product is damaged. I had to return it.
26,https://www.daraz.com.ng/products/8848-men-flaese-jacket-kp105757-8a-1103484168-c1024226058.html?search=1,Very disappointed. Jacket was discolored ,2.0
27,https://www.daraz.com.ng/products/8848-men-flaese-jacket-kp105757-8a-1103484168-c1024226058.html?search=1,It was quite perfect and also looks good,5.1

```

Figure 19 : Training Dataset

```

it best thing is low price it's so awesome prediction= {9.75963046}
Good product according to price.It is not best for gaming but it looks standard in daily usage basis for normal task. prediction= {9.81437324}
ham,saena hai keyboard light hales tatya rati keys haru visible hudeina ani tai mathi 2 ta key na fault sa nuta dabada arhi type hunna mouse not
lovely product at low price, truly value for money combo offer. absolutely love it and highly recommendable prediction= {11.96693915}
but hales haina keyboard raaro sa according to price but mouse all naramro sa .sender naramro sa prediction= {0.88886884}
it best thing is low price it's so awesome prediction= {9.75963046}
Keyboard na light na tara macbook sa light on haina k game prediction= {1.38856402}
KEYBOARD IS AWESOME BUT MOUSE SENSOR IS NOT WORKING prediction= [-1.87861421]
this keyboard 95% lights not working on my pc prediction= {0.83952433}
Keyboard is Ok But the Mouse very Bad . prediction= {1.2833445}
I got a problem with the mouse. prediction= {2.33531713}
my mouse is not working prediction= {1.39542269}
I loved it prediction= {1.41165638}
Batti halesa prediction= {1.39288323}
keyboard is not working prediction= [-1.13885562]
Satisfy prediction= {1.39288323}
The Accuracy is..... 68.18146679116889
Predictions using naive bayes:
[["Good product according to price.It is not best for gaming but it looks standard in daily usage basis for normal task.", 1], [{"ham,saena hai k
.....
{0, 1}]]

```

Figure 20 : Test Dataset